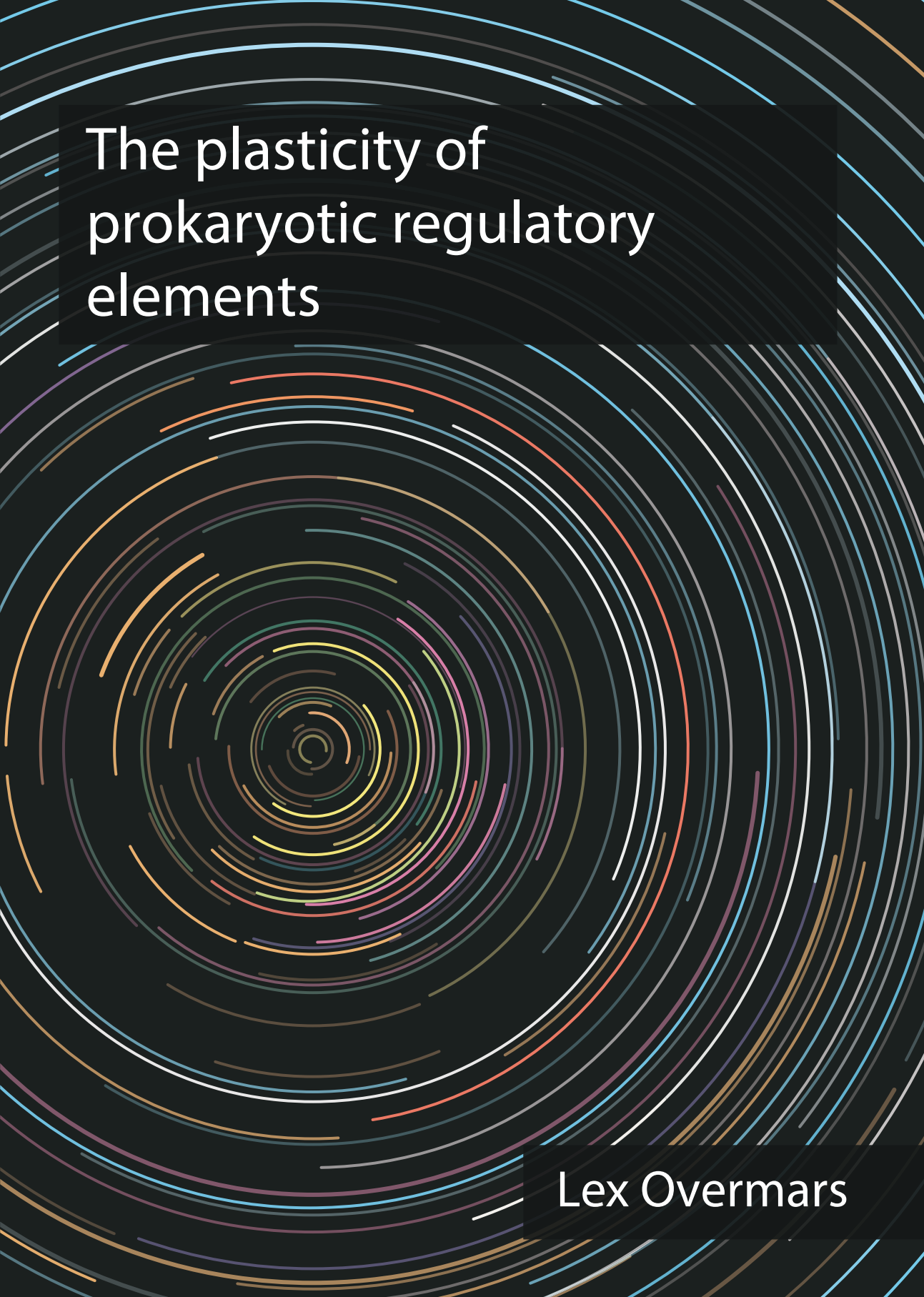
The plasticity of prokaryotic regulatory elements



Lex Overmars

ISBN: 978-94-028-0636-6

The research presented in this thesis was conducted at the Bacterial Genomics group at the Center for Molecular and Biomolecular Informatics, Radboud University Medical Center, Nijmegen, The Netherlands.
Printing of this thesis was financially supported by the Radboud University Nijmegen, the Netherlands.

"The plasticity of prokaryotic regulatory elements"

Printing: Ipskamp Printing, The Netherlands
Cover: Lex Overmars
Funding: This work was part of the BioRange programme of the Netherlands Bioinformatics Centre (NBIC), which was supported by the Netherlands Genomics Initiative (NGI).

The plasticity of prokaryotic regulatory elements

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken,
volgens besluit van het college van decanen
in het openbaar te verdedigen op

woensdag 31 mei 2017
om 16.30 uur precies

door
Lex Overmars
geboren op 12 maart 1985
te Voorst

Promotor:

Prof. dr. R.J. Siezen

Copromotoren:

Dr. C. Francke (Hogeschool Arnhem & Nijmegen)
Dr. S.A.F.T. van Hijum

Manuscriptcommissie:

Prof. dr. J.A. Veltman (voorzitter)
Prof. dr. H.J.M. Op den Camp
Prof. dr. ir. D. de Ridder (Wageningen University & Research)

Table of Contents

Chapter 1

General Introduction

1 Functional hierarchy in a prokaryotic cell: a bioinformatics perspective

The various roles of the DNA sequence

Prokaryotes have a remarkable ability to survive in nearly all environments on Earth. The survival skills are in essence encoded in a single DNA molecule (sometimes more), the chromosome. The chromosome embodies the genetic information and is located in the inside of the cell (Kleckner et al., 2014; Kuhlman and Cox, 2012; Robinow and Kellenberger, 1994). It is surrounded by the cytoplasm, where the cellular reactions take place, and one or two membranes (Gram-positive or Gram-negative bacteria, respectively). The relation between encoded information and the cellular chemistry and physics is maintained by a hierarchy of connected processes such as replication (DNA→DNA), transcription (DNA→RNA), translation (mRNA→protein) and catalysis (protein/RNA→increased reaction rates). At all levels of the hierarchy these processes are subject to regulation (i.e. control), which ultimately also is encoded in the DNA molecule(s). The term 'genome' is used to refer to the genetic material from the perspective of the conversion of the molecularly stored information to the whole of the functioning organism.

The majority of a prokaryotic chromosome consists of coding sequences in the form of genes. Genes are transcribed to various types of RNA, including mRNA, which is translated to protein, and tRNA and rRNA, which are involved in the translation process. In addition, several forms of RNA are encoded that relate to control, such as for instance small regulatory RNA, ncRNA, miRNA, asRNA and siRNA (Dinger et al., 2011; Papenfort and Vogel, 2010). Generally the DNA is densely coded, with approximately one gene per kilo-base of DNA (Bentley and Parkhill, 2004; Koonin and Wolf, 2008). Nonetheless, coding sequences are not the only functionally important elements of the prokaryotic chromosome. The chromosome also contains a large number of small recurring sequence elements. These elements may for instance be associated to protein or RNA binding sites, DNA or RNA structure, and involved in the maintenance of chromosome plasticity. Moreover, for various recurring sequences the function has not been clarified yet.

Arguably, the most-studied recurring DNA sequence elements are transcription factor binding sites, which have a characteristic composition that depends on the related transcription factor. Other well-studied binding DNA sequences are those of the promoter, where the process of transcription is initiated by recruitment of the RNA polymerase complex. In prokaryotes, the promoter region typically consists of two short characteristic sequences

that are separated by a defined number of nucleotides. The specific core promoter sequence is recognized by sigma factors, which are proteins that assist the RNA polymerase in binding to the promoter region (Feklístov et al., 2014; Gruber and Gross, 2003; Paget, 2015). Transcription is terminated at yet another recurring sequence element, rho-independent terminators, which consist of a 30-40 bp sequence that forms a stem-loop structure in the mRNA molecule to terminate transcription (Carafa et al., 1990). In the process of translation the prokaryotic ribosome binds to a sequence element called the Shine-Dalgarno sequence, which is located 5-13 nucleotides upstream of the start codon (Ma et al., 2002).

Characteristic recurring DNA sequences are also involved in other processes, such as chromosome replication and repair, genome plasticity and regulation of translation. Elements in the leader regions of mRNAs can impose a secondary structure that has a direct effect on translation initiation of the downstream coding sequences (Breaker, 2012; Marzi et al., 2008). These so called RNA switches or riboswitches can temporally regulate translation, leading either to repression or to activation of protein synthesis. Insertion Sequences (ISs) are also highly abundant in prokaryotic chromosomes and important for their plasticity. Typical ISs are between 0.7 and 3.5 kb in length, include one or two open reading frames (ORFs) which occupy the complete IS and terminate on both sides in a flanking imperfect terminal repeat sequence (IR) (Mahillon and Chandler 1998; Siguier et al., 2015). IS populations are capable of expanding within genomes and also participate in genome streamlining or trimming by facilitating DNA deletions (Siguier et al., 2014).

Sequence heterogeneity

Any DNA sequence that is associated with a particular function, be it coding for a particular protein or RNA molecule, a particular binding-site, or otherwise, could potentially allow some sequence heterogeneity without impairing the function. The impact of such a 'degeneracy' of the DNA element depends on the type of sequence and its role in the physiology of the cell. Sequence variation in a coding sequence can obviously affect the function of its corresponding product. However, within a coding sequence the consequence of a sequence change will highly depend on its position. For instance, the use of the three-nucleotide (triplet) genetic code in conjunction with the 20 amino-acid 'alphabet' imposes that several distinct codons ('synonymous' codons) are translated into the same amino acid. As a result, in the case of a gene, the coding DNA sequence may be different although the protein product is identical. Synonymous codons vary mostly at the third position

of the triplet. Synonymous codons are used with different frequencies, a phenomenon known as codon usage bias. The codon bias can be related to the G+C content, which varies over a wide range between prokaryotic species (Muto and Osawa, 1987), but also to control of the translation rate, in which it corresponds to tRNA abundance (Dittmar et al., 2005; Elf et al., 2003). Moreover, nucleotide changes that do impose a change in protein sequence in most cases do not affect protein function unless the affected amino acid is essential to that function.

Sequence degeneracy of binding sites can vary greatly. On the one hand the sequence may be tightly constrained like in the case of the binding sites of Type II restriction enzymes, where the sites are highly conserved. Already a slight deviation from the consensus binding sequence could lead to irreversible damage to the chromosome. The Type II restriction enzymes are part of a prokaryotic immune system designed to cut up viral DNA from infecting phages (D'haeseleer, 2006). On the other hand, the binding sequence related to some other DNA binding proteins appears far less constrained. The promoter sequence, for instance, contains a so-called 'TATAAT box' which is centered around 10 bp upstream of the transcription initiation site, and a sigma factor specific sequence (Feklístov et al., 2014; Paget, 2015) centered around -35 bp. In the case of the 'household' sigma-factor 70, which is present in all prokaryotes (Gruber and Gross, 2003; Paget and Helmann, 2003), the consensus motif is TTGACA. Nevertheless, the degree of conservation at each position of the promoter sequence ranges from 54% to 82% for each base in *Escherichia coli* (D'haeseleer, 2006). It is actually rare to find a promoter that matches the consensus sequence exactly, with most promoters matching only 7–9 out of the 12 bases. In the case of transcription factor binding sites a slight degeneracy or deviation from the consensus sequence is actually needed for an optimal function. It was shown that a perfectly symmetrical binding sequence caused tight binding of the LacI repressor even when not induced, thereby impairing the regulatory function (Sadler et al., 1983). Differences in binding sequence degeneracy have been proposed to be related to a difference in global and local regulation, where global regulators are associated to less conserved binding sites (Francke et al., 2008).

The sequence degeneracy of structural elements such as terminators and riboswitches has a different impact on their function and binding affinities of proteins that target those elements. The sequence that makes up the structural characteristic of the element is relatively less prone to sequence variation. The sequence conversation of the binding part of the element (if present) is again subject to the specificity of the target proteins. For example, an *in silico* analysis of t-box elements, a specific type of anti-termination

element, showed that only smaller regions of these elements were highly conserved (Wels et al., 2008)

As a consequence of the fact that the conservation of the structural features is more important than the precise composition of the primary sequence, other approaches might be required when identifying these elements. For example, a combination of secondary structure alignment and sequence homology searches was applied to identify possible new riboswitches (Weinberg et al., 2010). Also rho-independent terminators can be relatively straightforwardly characterized by their structural feature; a short GC-rich stem-loop followed by a chain of Uracil residues, but are difficult to identify using sequence conservation only (Kingsford et al., 2007) .

Global and local structure of DNA

Unlike the 'linear' DNA of most eukaryotes, prokaryotic chromosomes are typically circular. When the DNA helix has the normal number of base pairs per helical turn (about 10.5 base pairs) it is in the relaxed state. If the helix is overwound so that it becomes tighter, the edges of the narrow groove move closer together. If the helix is underwound, the edges of the narrow groove move further apart. DNA supercoiling refers to this over- or under-winding of a DNA strand. Overwound DNA exhibits positive supercoiling, whereas underwound DNA exhibits negative supercoiling. The DNA helicity in a cell is determined by the relative activity of DNA gyrase, the enzyme that introduces negative supercoils, and topoI and topoIV, two enzymes that relax the supercoils (Menzel and Gellert, 1983; Zechiedrich et al., 2000). Chromosomes are exposed to many biochemical reactions and biomechanical processes that require specific types of DNA movement. Local negative supercoils enable the transition from duplex- to single-stranded conformation, in which DNA replication, transcription and recombination can occur (Nöllmann et al., 2007). To accommodate these processes different levels of DNA organization exist within the chromosome. In the case of *Escherichia coli*, the genome is organized in four individual macrodomains (labeled Ter, Ori, Right, and Left) and two less-structured regions. The macrodomains are precisely localized within the cell throughout the cell cycle and are associated with specific binding proteins (Dame et al., 2011). At a smaller scale of ~10kb, topological domains can be discerned, which are formed by supercoiled structures (Deng et al., 2005). The barriers of those domains are not positioned at fixed locations but seem more randomly distributed. There are even smaller loops of a few hundred base pairs made by specific transcription factors. Experimental data show a correlation between transcriptional activity and the number and stability of looped DNA domains in a particular region (Postow et al.,

2004; Stein et al., 2005). This is also dependent on the physiological state of the cell; fast-growing bacteria have more transcriptional activity, mostly at genes coding for ribosome components and other parts of the translation machinery, and have more looped DNA domains (Dillon and Dorman, 2010)

The importance of BioIT

Classically, functional DNA sequences were found and characterized experimentally. For instance, transcription factor binding sites were determined using experimental assays such as DNase footprinting and reporter constructs (D'haeseleer, 2006). The availability of whole genome sequences has led to experimental advances. Chromatin immune-precipitation coupled to tiled microarrays (ChIP-seq) is used to examine genome-wide binding of transcription factors. In high-throughput SELEX (Systematic Evolution of Ligands by EXponential enrichment), binding specificity is determined by allowing a protein to select its target sites from a pool of DNA strands containing randomized sequences (Jolma et al., 2013). Nevertheless, experimental procedures to determine the exact binding sites are expensive and time-consuming. Nowadays, the availability of whole genome sequences has enabled computational approaches to search for overrepresented or conserved DNA sequences upstream of functionally related genes (i.e. co-expression, or similar function annotation) (D'haeseleer, 2006). The huge advantage of being able to use computational approaches is also reflected in the number of methods that have been developed for the analysis and discovery of binding sites of transcriptional regulators; over 200 tools and methods were developed in 2010 (Ladunga, 2010).

The advancements in both computational and experimental methods have led to the definition of more sequence motifs. The abundance of these derived sequence motifs and their increasing utilization in the reconstruction of regulatory networks and the identification of the regulatory control of individual genes makes them an essential tool in this post-genomic era (D'haeseleer, 2006).

Genomic elements

Coding elements

The density of protein-coding genes is high in prokaryotic genomes and the number of genes is strongly correlated to the genome size. The intergenic distances for the average genes are small. The distribution of these distances is bimodal, with the first peak, at ~0 bp, corresponding to the densely organized genome segments, primarily within operons, and the second peak,

at ~100bp, corresponding to inter-operonic regions (Koonin and Wolf, 2008).

Transcription regulation and transcription binding factors

Transcription regulation (gene expression) in a prokaryotic cell plays a crucial role in the long-term cellular response to changes in the internal and external environment (Bauer et al., 2010). Transcription factors can either be regulatory RNA molecules or DNA-binding proteins. In general the total response is built up from the response of one or a few so-called transcription units (TU). These units are ordered DNA segments consisting of the following components: a regulatory region, a transcription start site, one or more ORFs and a transcription termination site (Fig. 1).

The regulatory region contains the promoter and transcription factor binding site(s), where transcription factors (TFs) can bind to control the recruitment of the RNA polymerase. In the case of protein transcription factors at least two domains are involved, one domain that serves as a sensor for the stimulus (e.g. via protein-protein interaction or binding of a ligand) and a responsive domain that interacts with the target DNA binding-sequence (Jacob, 1970). Intracellular stimuli (i.e. signals) are mostly handled by one-protein transcription factors that carry both domains (Ulrich et al., 2005). In some cases the protein is actually a bifunctional enzyme (Commichau and

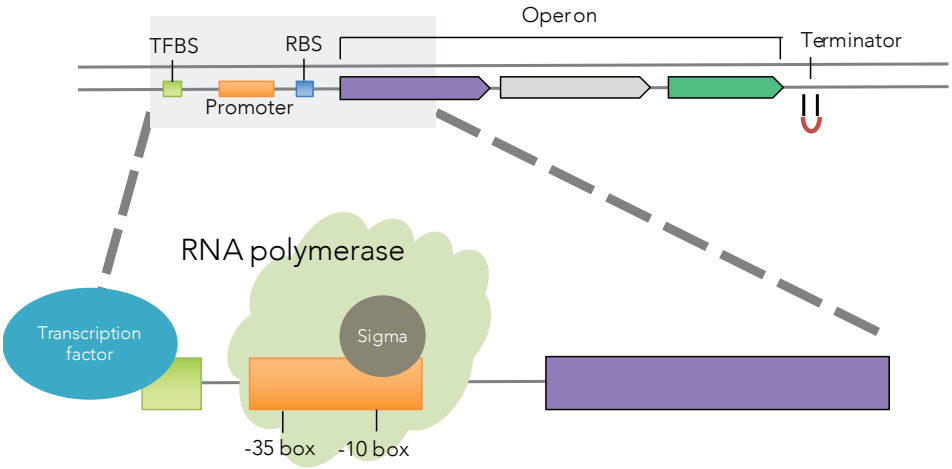


Figure 1. Schematic view of a transcription unit in a prokaryotic genome. Upstream of the coding genes is a promoter sequence located which provides a site for RNA polymerase to bind and initiate transcription. Genes that are located in the transcription unit are transcribed into one mRNA molecule and are denoted as an operon. Transcription of an operon can be activated or repressed when a transcription factor (TF) binds to its corresponding transcription factor binding site (TFBS) in the promoter region upstream of the operon. A transcription terminator marks the end of the operon during transcription. The ribosomal binding site (RBS) is transcribed but not translated; it is bound by the ribosome when initiating protein translation.

Stülke, 2008). Extracellular stimuli that cannot easily pass the membrane barrier are handled by two-component systems (Stock et al., 2000). These systems consist of a membrane bound sensor protein (histidine kinase) that dimerizes and auto-phosphorylates upon signal binding, and a cytoplasmic protein (response regulator) that receives the phosphoryl-group and then can bind to the appropriated binding sites on the DNA. Extracellular stimuli can also trigger specific ECF sigma-factors (Campagne et al., 2015).

A TF commonly regulates various genes whose 'upstream' transcription factor binding sites have comparable but not identical sequences. We can distinguish TFs that increase the rate of transcription (i.e. up regulation), which are denoted as activators and TFs that decrease the rate of transcription (i.e. down regulation), which are denoted as repressors. The binding of a repressor to the promoter region, thereby interfering with the binding of the RNA polymerase or its procession, is the most common mechanism of down regulation. Activators generally bind just upstream of the promoter and thereby recruit the RNA polymerase directly. The control of TFs over the transcription activity for a set of genes, also denoted as target genes, is commonly reconstructed in a transcriptional regulatory network (TRN). The collections of genes that are controlled by the same TF compose a regulon.

Previously identified regulatory DNA motifs such as transcription factor binding sites and their corresponding regulons have been collected in different resources. Some of these resources specifically aim at single organisms, such as RegulonDB, which is developed for *E. coli*, while others aim at exclusively prokaryotes or eukaryotes (e.g. Jaspar (Mathelier et al., 2013)). Moreover, the level of manual curation differs between the different resources. An overview of the most used and/or most relevant resources for bacterial genome analysis are given in table 1.

Table 1. Prokaryotic regulon databases and resources.

Name	Url	Reference
PePPER	http://pepper.molgenrug.nl	Jong et al., 2012
DBTBS	http://dbtbs.hgc.jp	Sierro et al., 2008
RegPrecise	http://regprecise.lbl.gov	Novichkov et al., 2013
FITBAR	http://archaea.u-psud.fr/fitbar	Oberto, 2010
RegTransBase	http://regtransbase.lbl.gov	Cipriano et al., 2013
JASPAR	http://jaspar.genereg.net	Mathelier et al., 2013

The molecular interactions involved in DNA binding of TFs

Site-specific DNA binding by proteins is a feature of the regulatory processes that maintain, expand, and express genetic information such as replication, recombination, transposition, and transcription. The chemical and physical mechanisms that underlie sequence-specific recognition of regulatory elements by DNA-binding proteins can be divided in two classes: direct- and indirect- readout (Michael Gromiha et al., 2004). Direct readout refers to the hydrogen bonds between proteins and the unique extra-cyclic substituents at C-4 of pyrimidines, and C-6 and N-7 of purines. These groups provide a base pair-specific pattern of hydrogen bond donors and acceptors in the major groove of DNA that can be directly read by a complementary pattern of amino acid side chain donors and acceptors (Aeling et al., 2006). In the case of indirect readout, a TF recognizes aspects of the DNA structure such as intrinsic curvature, topology of major and minor grooves, ordered water structures, local geometry of backbone phosphates, and flexibility or deformability (Michael Gromiha et al., 2004).

Structural elements

Secondary structure elements both on the DNA and RNA are of great functional importance. DNA structural elements, such as cruciforms, are important for the chromosomal structure and have thereby an important function for regulating biological processes (Brázda et al., 2011). RNA structural elements are important in various processes such as transcription (i.e. rho-independent terminator stem-loops) and translation (i.e. tRNA cloverleafs). Structured RNAs can be detected on the DNA by comparative genomics, in which homologous sequences can be identified and inspected for mutations that conserve RNA secondary structure (Weinberg et al., 2010).

Prokaryotes use two main modes of terminating transcription: rho-independent termination (also denoted as intrinsic termination) and rho-dependent termination. Rho-dependent terminators are inconsistent in terms of sequence similarity but all require the protein factor rho for termination. At the DNA level, rho-independent terminators generally consist of a GC-rich palindrome followed by a run of T's defining the site of termination. At the RNA level this give rise to a stem-loop structure and a run of U's, which induce RNA polymerase to pause, destabilize and disengage for releasing RNA without the involvement of any proteins (Naville and Gautheret, 2010a). In addition, prokaryotes developed a markedly wide range of termination systems, each of them coupling formation of a transcription termination structure to some sensing mechanism able to detect a determining environmental factor or

event. The type of signal detected is commonly used to classify attenuators into major families: riboswitches bind small metabolites, T-boxes bind tRNAs, and other types respond to protein factors or temperature (Naville and Gautheret, 2010b).
Riboswitches are mRNA sequences that contain specific ligand-binding (sensor) domains along with a variable sequence, termed the expression platform, which enables regulation of the downstream coding sequences. Metabolites present in cells above threshold concentration can be directly sensed and specifically bound by the sensor domains, which induce a conformational change in the expression platform, which in turn leads to modulation of downstream events (Serganov and Nudler, 2013). Riboswitches are capable of regulating gene expression by directly sensing a metabolite without requiring any intervention of proteins (Winkler et al., 2002). They have been shown to be involved in the control of different biosynthetic processes such as thiamine, riboflavin, cobalamine, adenine, guanine and lysine biosynthesis (Mandal and Breaker, 2004). The selectivity of riboswitches is entirely encoded in their conserved sensing domains. These recognition sites vary greatly in the size and complexity of their secondary and tertiary structures (Serganov and Nudler, 2013).

The T-box system represents a special class of riboswitch RNAs in which binding of a specific uncharged tRNA to the 5′ region of the transcript, without any required factor, promotes expression of the downstream genes, which usually encode the cognate aminoacyl-tRNA synthetase, or the proteins that synthesize or transport the cognate amino acid.(Vitreschak et al., 2003, 2008; Wels et al., 2008). A 30-nt motif that is well-conserved and positioned in the 3'-region of the terminator/anti-terminator loop can be used to identify T-boxes *in silico* (Wels et al., 2008). Comparative sequence analysis of the elements suggests they can evolve separately from the genes they control (Vitreschak et al., 2008; Wels et al., 2008)

Replication and recombination

DNA Replication in prokaryotes is orchestrated by a large number of proteins and enzymes, which in turn recognize and bind different DNA sequences. The replication machinery starts at a specific sequence called the origin of replication, or ori. Most prokaryotes have one origin of replication on their chromosome; in *E. coli* the corresponding sequence is AT-rich (as it is in most organisms) and about 245 nucleotides long. DnaA boxes are non-palindromic sequences of 9 nucleotides, which are clustered close to the ori. Replication begins with the binding of DnaA to these DnaA boxes, which leads to strand separation in the AT-rich region within the ori. DnaA and DnaA boxes are

well conserved across bacterial species and are abundant in many bacterial genomes (Touzain et al., 2011). The DnaA box motif consensus sequence in the *E. coli* genome is TTATNCACA, where it occurs 107 times.

Repair of DNA double-strand breaks by homologous recombination is crucial to maintain functional genomes. The major pathway of DNA break repair in *E. coli* requires RecBCD. The RecBCD enzyme is a large complex of three polypeptides with both DNA-unwinding (helicase) and DNA hydrolysis (nuclease) activity. Beginning at a DNA double-strand end, it unwinds DNA and, when it encounters a short sequence called Chi, makes a new 3′ end at which it begins loading RecA proteins onto the single-stranded DNA (ssDNA) generated by continued unwinding (Smith, 2012). RecA then promotes the exchange of this ssDNA for its equal in an intact homologous DNA molecule. The sequence of these Chi sites is different in different organisms, in *E. coli* the sequence is 5'-GCTGGTGG-3' and occurs about a 1000 times in the genome (El Karoui et al., 1999).

Repetitive sequences

DNA repeats are a source of genome plasticity, sequence recombination and functional overlap. They are targeted by recombination processes and thereby responsible for duplications, deletions and rearrangements of the DNA. Repeats can include complete genes or operons, which can lead to functional redundancy when amplified within a genome. Genes can also be co-expressed when similar copies of a repeat in their regulatory region are present (Treangen et al., 2009a). However, questions concerning various of the different interspersed sequences have remained unanswered. For many families of these sequence elements, the reason behind their specific location distribution or their role in chromosome organization are still a puzzle.
The proliferation of repetitive or selfish elements in genomes is accountable for many of the DNA repeats found in prokaryotic genomes (Treangen et al., 2009a). Transposable elements (TEs) are the best-studied and the most abundant 'selfish' elements. Two groups have been distinguished based on their structure and transposition mechanism: i) class I elements (also denoted as retrotransposons), and ii) class II elements (also denoted as DNA transposons). Class I elements enable transposition using an RNA intermediate, whereas the highly abundant class II elements use a DNA intermediate. The most abundant members of the class II family of elements are ISs (insertion sequences elements). The length of these mobile genetic elements ranges between 0.7 and 3.5 kbp. Typically, they carry one or two ORFs (the transposase) and are flanked by imperfect terminal repeat sequences (Mahillon and Chandler, 1998).

Other families of DNA repeats have been identified in many prokaryotic genomes, but their origin and exact function have not been established (Treangen et al., 2009a). Many of such elements are found widespread throughout bacterial and archaeal phyla, whereas others appear to be species-specific, or even subspecies-specific. A class of well-characterized repetitive sequences is that of the CRISPRs. CRISPRs are found in many bacterial and archaeal genomes (Jansen et al., 2002). They are composed of short repeats (24 to 47 nt), which are separated by 'spacers' of similar length (Barrangou et al., 2007; Horvath 2013). The spacer sequences are highly similar to sequences of phages and provide, by the use of a mechanism comparable to RNAi in eukaryotic organisms, resistance to foreign genetic elements. Another repetitive element, which is found in *gammaproteobacteria* is the Repetitive Extragenic Palindromic sequence (REP) (Bachellier et al., 1999). REPS are well conserved within each genome. So far, the function of these elements has remained unresolved, though the element was discovered more than 30 years ago (Higgins et al., 1982). The REP sequences are palindromic and their length ranges between 21 and 65 nt. In *E. coli* REPs occupy a large fraction of the total intergenic space (Tobes and Ramos, 2005).

Database resources are available for specific types of repetitive sequences (Table 2). ISfinder is a dedicated database of previously identified bacterial insertion sequences (ISs) (Siguier et al., 2006) whereas CRISPdb is a resource in which microbial genomes have been pre-processed in search for CRISPRs structures (Grissa et al., 2007a).

Table 2. Publicly available database resources with previously identified repetitive sequences.

Name	Element	Url	Reference
CRISPRdb	CRISPRs	http://crispr.u-psud.fr/	Grissa et al., 2007a
ISfinder	Insertion sequences	http://www-is.biotoul.fr/	Siguier et al., 2006
PROPHAGEdB	Prophages	http://bioserver1.physics.iisc.ernet.in/prophagedb/	Srividhya et al., 2007
Microorganisms Tandem Repeats Database	Tandem repeats	http://tandemrepeat.u-psud.fr/	Denœud and Vergnaud, 2004

Strategies for the identification of DNA elements

The sequence of a short functional DNA sequence is often described in terms of 'a motif', which is a sort of sequence average for the particular functional DNA sequence in a genome or across genomes. The term 'consensus sequence' is also often used and refers to the predominant nucleotide at each position of the element. Computer-aided identification of particular DNA elements in

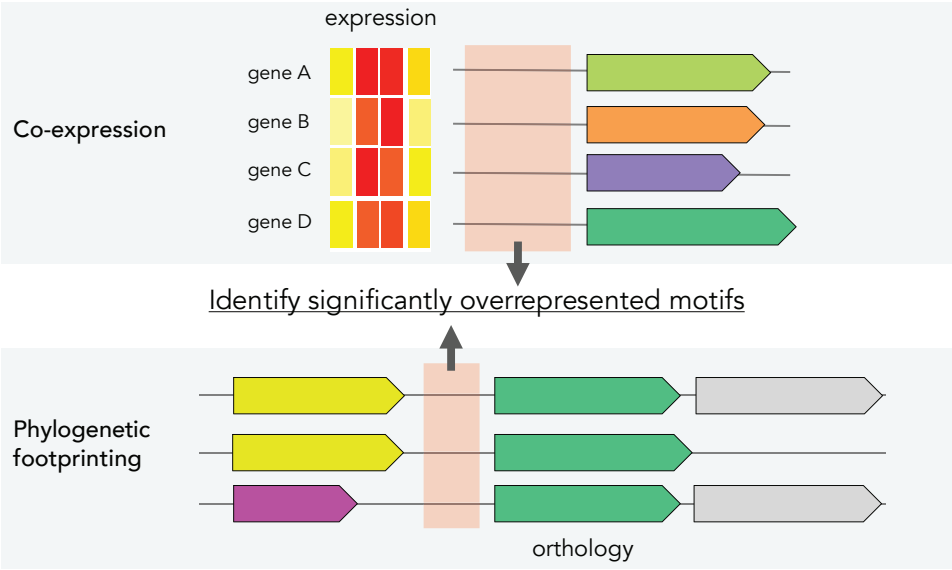


Figure 2. Unsupervised approaches for finding regulatory sequences. The upstream regions can be searched for significantly overrepresented sequences.

a prokaryotic genome is commonly achieved using three different strategies (Mrázek, 2009): i) unsupervised identification, where a set of DNA regions is selected that presumably contain a particular sequence element and these regions are evaluated for statistically over-represented motifs; ii) supervised identification, where a set of sequences representing a particular element is used to define a sequence motif that can be used, to find similar sequences in the genome; and iii) exploratory identification, where a genome is searched for sequence elements whose genomic distribution deviates in a statistical sense.

Unsupervised identification

Unsupervised identification approaches are used when the expected location of the sequence element can be established but no clear image exists of the constituting nucleotides. A common approach is the search for significantly

Table 3. Sequence motif discovery tools

Name	Url	Reference
MEME	http://meme-suite.org/	Bailey et al., 2015
PhyloGibbs	http://www.phylogibbs.unibas.ch/	Siddharthan et al., 2005
GLAM2	http://meme-suite.org/	Frith et al., 2008
Clover	http://zlab.bu.edu/clover/	Frith et al., 2004
RSAT tools	http://embnet.ccg.unam.mx/rsa-tools/	Thomas-chollier et al., 2011

overrepresented sequence motifs in the upstream region of a set of genes or a cluster of genes that experimentally was found to be co-expressed and hence, might be co-regulated (Fig. 2) (Conlon et al., 2003; Wels et al., 2011). Popular tools for unsupervised identification include MEME (Bailey et al., 2015), Phylogibbs (Siddharthan et al., 2005) and GLAM2 (Frith et al., 2008) (see Table 3). These tools typically require a set of DNA regions in which one expects one or multiple conserved DNA sequence elements.
A different approach in which unsupervised identification (but also semi-supervised identification) is applied is called phylogenetic footprinting (Fig. 2) (Francke et al., 2011; McCue et al., 2001; Wels et al., 2006). Phylogenetic footprinting relies on the fact that the evolutionary conservation of regulatory sequences is higher than the overall conservation of the intergenic DNA sequence. The typical workflow can be described as follows: i) selection of a gene of interest, ii) retrieval of its orthologs in selected other species, iii) extraction of their (potential-) regulatory regions and iv) analysis of these regions using an unsupervised motif finding program. However, some factors can complicate the analysis. First, it can be difficult to determine which region is the possible regulatory region. Moreover, selecting orthologs to include can be tricky; genomes should differ enough to have differences in their intergenic regions, but should be close enough to still have sequence similarity in the regulatory region.

Supervised identification

Supervised identification is applied when a sequence element is known and/or has been defined and one wants to find all similar sequences in a genome or multiple genomes. An adequate description of the sequence of the element is a critical step to enable the retrieval of all potential positions in a genome (Bailey, 2008; Frith et al., 2008; Liu et al., 2012). Perhaps, the most simplistic approach is to determine a consensus sequence (Fig. 3B) and consider every sequence that matches this sequence to be relevant. The approach is suitable for sequences that are highly conserved but less so for more degenerate elements. Therefore a different approach is required for these types of elements.

A more sophisticated way to represent a DNA motif is through a Position Frequency Matrix (PFM). Instead of only keeping track of the consensus sequences, the frequency of each base occurring at a certain position in the sequences that build the motif is reported (Fig. 3C). Frequency matrices can be visualized in a 'sequence logo' (Crooks et al., 2004), in which the DNA motif is visualized by scaling the bases based on their observed frequencies (Fig. 3D).

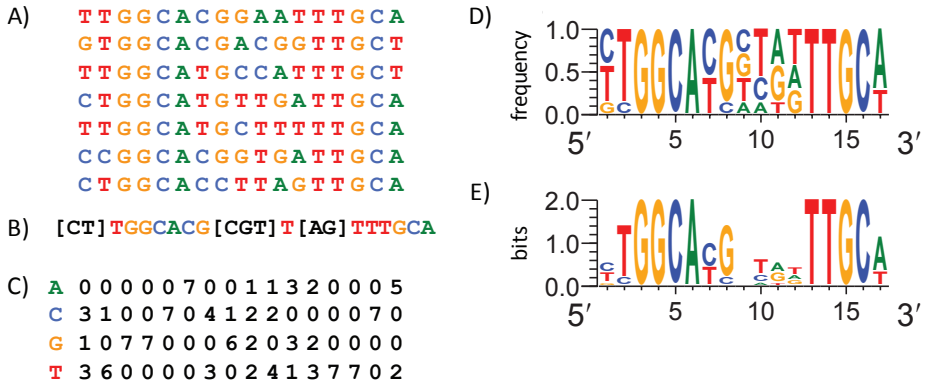


Figure 3. Sigma54 promoter sequences in *Lactobacillus plantarum* WCFS1. A) The sequences of 7 predicted sigma 54 promoter sites (Francke et al., 2011) in *Lactobacillus. plantarum* WCFS1 **B)** Degenerate consensus sequence of those 7 sequences. **C)** Frequencies of nucleotides at each position, **D)** sequence logo representation of the observed frequencies at each position and **E)** sequence logo representation of the information content in the sequence motif.

PFMs are commonly converted to a Position-specific score matrix (PSSM; also denoted as Position Weight Matrix (PWN) or position-specific weight matrix (PSWM)). A PSSM includes not only observed frequencies but in addition also takes in to account the background probability of each base. The first step to create a PSSM from a PWN is to divide the nucleotide frequencies at each position by the number of sequences, thereby normalizing the frequencies. The resulting matrix is called a position probability matrix (PPM). In order to avoid 0 values in the matrix, pseudocounts are commonly applied when calculating PPMs. By using a pseudocount, a small number is added to each value in the matrix. The selection of pseudocount values is often purely empirical. The elements in the PSSM are then commonly calculated as log likelihoods. This could be the final step in the process of creating a PSSM. However, when one is searching for a DNA motif in a high GC% genome, the probability of encountering a G or C is higher. In that case it could be argued that a conserved A or T should be assigned higher significance. To account for genomic differences, a probabilistic model of the sequence background is added. In a 60% GC-content genome the applied background probabilities for A, T, G, C are 0.3, 0.3, 0.2 and 0.2. This means that when one applies this background model the scores not only represent the similarity to the motif model, but also the probability of encountering the sequence in its sequence background.

PSSMs are applied in many tools that are used to search for sequence motifs, such as the popular tool MEME (Bailey et al., 2015). Nevertheless, tools making use of position-specific weight matrices provide scores and associated

rankings that reflect probability rather than similarity. As the number of sequences on which a PSSM is based is limited, the desired sampling of the query sequence is not provided. As a consequence, probabilistic scores will be skewed. It has been shown that when applying PSSMs to search genomic sequences, the (uniform-) correction for the genomic background can lead to overestimation of the importance of the different bases within a motif, due to the uneven distribution of k-mers in real genomes, leading to larger numbers of false positives (Erill and O'Neill, 2009). In addition, introducing additional parameters to the score such as the pseudocount –to ensure no 0 values are encountered-, perhaps deviate the model from the DNA motif. To overcome these issues we formulated a similarity search method (Francke et al., 2011). Given any number of input sequences of size i, the nucleotide frequency fN(j) (where N ∈ A,C,T,G; and frequency is in terms of fraction) at every position j = 1 to i can be used directly to provide all target sequences of size i with a score by just adding up the input-based frequencies that relate to the nucleotide composition of the target. Division of the score by the length of the sequence i results in a 'similarity' score that can range from 0 to 1. Dividing this number by the highest attainable score given the input matrix then yields a relative 'similarity' score. The method was tested and appeared at least as good to identify putative regulatory elements on basis of known input motifs as the commonly used tools, yet providing a similarity score that is far easier to interpret and use (Francke et al., 2011).

Exploratory identification

The identification of sequence motifs that are atypical or deviating in statistical terms is denoted as exploratory motif finding. The 'deviation' that is mostly searched for relates to abundance of a sequence at significantly higher frequencies than expected, or significantly lower frequencies than expected. These types of analysis are typically done to identify short repetitive equences and various tools have been developed to accommodate them. An example is AIMIE (Mrázek et al., 2008), which can be used to identify significantly overrepresented short motifs such as repeated extragenic palindrome (REP) elements and CRISPR repeats. CRISPRFinder is another tool specifically developed to identify CRISPR sequences (Grissa et al., 2007b). Numerous tools and algorithms are publicly available for *de novo* identification of repeat families, such as the RepeatScout algorithm (Price et al., 2005), Repeatoire (Treangen et al., 2009b) and RepeatFinder (Volfovsky et al., 2001).

Prediction of translation initiation

An important prerequisite of efficient identification of true regulatory sequences is a correct annotation of coding sequences. Various computational methods have been developed to identify coding sequences from Open Reading Frames (ORFs) with low error rate. An overview of available tools and resources is given in Table 4. Nevertheless, the identification of the correct translation initiation site is still ambiguous. This is illustrated in a study by Dunbar and colleagues, which showed that only 53% of the orthologs among 5 Burkholderia genomes have consistently annotated translation initiation sites (TISs) in RefSeq. While incorrectly annotated genomes can flaw the identification of regulatory sequences, a quality measure for the annotation of TIS annotation does not exist yet. In this thesis we describe a novel measure for the genome-wide quality of TIS annotations.

Table 4. ORF and TIS prediction tools and resources.

Name	Type	Reference
Prodigal	*De novo* ORF prediction	Hyatt et al., 2010
Glimmer 3	*De novo* ORF prediction	Delcher et al., 2007
Easygene	*De novo* ORF prediction	Larsen and Krogh, 2003
Tico	Post-processing	Tech et al., 2005
GenePRIMP	Post-processing	Pati et al., 2010
ProTISA	Post-processing & resource	Hu et al., 2008

Analysis of DNA elements

The human mind may not intuitively understand complex statistical models but our brain is excellent at recognizing patterns from visual displays. For that reason, visualization can be a great asset to facilitate genome analysis. The integration of genome data, annotation data and (predicted-) regulatory elements can also specifically guide the identification of regulatory elements. We can distinguish three types of available approaches that include visualization and are beneficial for the analysis of (regulatory-) DNA elements (Table 5): i) generic genome browsers, ii) genome analysis platforms and iii) genomic context viewers. Widely used generic genome browsers such Artemis (Rutherford et al., 2000) and UCSC genome browser (Fujita et al., 2011) are capable of including generic genetic elements, such as regulatory elements. These types of genome browsers can be specifically useful to visualize experimental data with predicted DNA elements, as they can integrate multiple 'data-tracks'. Analysis platforms such as the JGI Genome portal (Nordberg et al., 2014), Microscope (Vallenet et al., 2013) and Microbes Online have not only a plethora of annotation data available but also offer useful (comparative-) genome visualizations. However, they

generally lack flexibility in adding user-defined DNA elements and flexible visual comparisons. The third type of view comparatively visualizes the genomic context of genes, often to address conservation of gene order between orthologous genes, also denoted as 'synteny'. Examples are PSAT (Fong et al., 2008) and Absynte (Despalins et al., 2011). In this thesis we describe two visualization tools, MGcV (Microbial Genomic context Viewer) and CiVi (Circular Visualization for Microbial Genome), that were specifically developed to enhance small-scale genome analysis such as the analysis of regulatory sequences.

Table 5. Visualization tools and resources that can aid the analysis of regulatory elements

Name	Type	Url	Reference
UCSC Genome Browser	Generic browser	http://genome.ucsc.edu	Rosenbloom et al., 2015
Artemis	Generic browser	http://www.sanger.ac.uk/Software/Artemis	Rutherford et al., 2000
JGI Genome portal	Analysis platform	http://genome.jgi.doe.gov/	Nordberg et al., 2014
MicroScope	Analysis platform	http://www.genoscope.cns.fr/agc/microscope	Vallenet et al., 2013
MicrobesOnline	Analysis platform	http://www.microbesonline.org	Dehal et al., 2010
GCView	Context viewer	http://toolkit.tuebingen.mpg.de/gcview	Grin and Linke, 2011
PSAT	Context viewer	http://www.nwrce.org/psat	Fong et al., 2008
Absynte	Context viewer	http://archaea.u-psud.fr/absynte	Despalins et al., 2011
MGcV	Context viewer	http://mgcv.cmbi.ru.nl	Overmars et al., 2013
CiVi	Circular visualization	http://civi.cmbi.ru.nl	Overmars et al., 2015

Outline of this thesis

The research presented in this thesis involves the functional analysis of regulatory sequences in prokaryotic genomes and the development of identification and visualization strategies that support the analysis. In **chapter 2** we describe two different web-applications that have been developed specifically to facilitate the analysis of regulatory elements and other functional DNA elements. McGV, described in **chapter 2a**, is a linear genomic context viewer with a very flexible and easy-to-use interface. In **chapter 2b** we describe CiVi, a circular viewer that can be used to create genome-wide visualizations. Both tools offer integration of identified (potential-) regulatory elements with existing annotation data and results of experiments, such as expression data.

While not only focusing on elements that are 'classically' considered to have a regulatory function, such as transcription factor binding sites, we investigated the presence of overrepresented repeated DNA sequences in **chapter 3**. We formulate a generalized strategy to characterize newly identified repeated DNA sequences based on i) the abundance and genome-wide distribution, ii) the taxonomic distribution and iii) the distribution with respect to the local gene organization.

Chapter 4 assigns a consistent putative biological role to a specific and very abundant repeated DNA element in *Escherichia coli*, known as Repetitive Extragenic Palindromic sequences (REPs). By analyzing the positional distribution within the genome and by analyzing the sequence characteristics and expression dynamics of adjacent genes we linked REP elements to the reduction of transcription interference caused by transcription induced supercoiling.

Incorrectly annotated translation initiation sites (TISs) can obscure the identification of true regulatory sequences. In **chapter 5**, we formulate a novel, reference-independent quality measure to score the TIS annotation in genomes. Our quality measure enabled us to evaluate the TIS annotation quality of publicly available prokaryotic genomes and analyze various factors that are believed to affect TIS annotation quality. In addition, we have developed a correction procedure for the annotation of microbial TISs that can supplement existing computational approaches.

Chapter 6 summarizes the main conclusions of this thesis and provides perspectives on further studies and future applications.

References

Aeling, K.A., Opel, M.L., Steffen, N.R., Tretyachenko-Ladokhina, V., Hatfield, G.W., Lathrop, R.H., and Senear, D.F. (2006). Indirect Recognition in Sequence-specific DNA Binding by *Escherichia coli* Integration Host Factor: the role of DNA deformation energy. *J. Biol. Chem.* **281,** 39236–39248.

Bachellier, S., Clément, J.-M., and Hofnung, M. (1999). Short palindromic repetitive DNA elements in enterobacteria: a survey. *Res. Microbiol.* **150**, 627–639.

Bailey, T. (2008). Discovering Sequence Motifs. *Comparative Genomics*, pp. 231–251.

Bailey, T.L., Johnson, J., Grant, C.E., and Noble, W.S. (2015). The MEME Suite. *Nucleic Acids Res*. **gkv416.**

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D.A., and Horvath, P. (2007). CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. *Science* **315**, 1709–1712.

Bauer, A.L., Hlavacek, W.S., Unkefer, P.J., and Mu, F. (2010). Using Sequence-Specific Chemical and Structural Properties of DNA to Predict Transcription Factor Binding Sites. *PLoS Comput Biol* **6**, e1001007.

Bentley, S.D., and Parkhill, J. (2004). Comparative Genomic Structure of Prokaryotes. *Annu. Rev. Genet.* **38**, 771–791.

Brázda, V., Laister, R.C., Jagelská, E.B., and Arrowsmith, C. (2011). Cruciform structures are a common DNA feature important for regulating biological processes. *BMC Mol. Biol*. **12**, 33.

Breaker, R.R. (2012). Riboswitches and the RNA world. *Cold Spring Harb. Perspect. Biol.* **4**.

Campagne, S., Allain, F.H.-T., and Vorholt, J.A. (2015). Extra Cytoplasmic Function sigma factors, recent structural insights into promoter recognition and regulation. *Curr. Opin. Struct. Biol.* **30**, 71–78.

Carafa, Y. d'Aubenton, Brody, E., and Thermes, C. (1990). Prediction of rho-independent *Escherichia coli* transcription terminators: A statistical analysis of their RNA stem-loop structures. *J. Mol. Biol.* **216**, 835–858.

Cipriano, M.J., Novichkov, P.N., Kazakov, A.E., Rodionov, D.A., Arkin, A.P., Gelfand, M.S., and Dubchak, I. (2013). RegTransBase – a database of regulatory sequences and interactions based on literature: a resource for investigating transcriptional regulation in prokaryotes. *BMC Genomics* **14**, 213.

Commichau, F.M., and Stülke, J. (2008). Trigger enzymes: bifunctional proteins active in metabolism and in controlling gene expression. *Mol. Microbiol.* **67**, 692–702.

Conlon, E.M., Liu, X.S., Lieb, J.D., and Liu, J.S. (2003). Integrating regulatory motif discovery and genome-wide expression analysis. *Proc. Natl. Acad. Sci.* **100**, 3339–3344.

Crooks, G.E., Hon, G., Chandonia, J.-M., and Brenner, S.E. (2004). WebLogo: A Sequence Logo Generator. *Genome Res*. **14**, 1188–1190.

Dame, R.T., Kalmykowa, O.J., and Grainger, D.C. (2011). Chromosomal macrodomains and associated proteins: implications for DNA organization and replication in gram negative bacteria. *PLoS Genet.* **7**, e1002123.

Dehal, P.S., Joachimiak, M.P., Price, M.N., Bates, J.T., Baumohl, J.K., Chivian, D., Friedland, G.D., Huang, K.H., Keller, K., Novichkov, P.S., et al. (2010). MicrobesOnline: an integrated portal for comparative and functional genomics. *Nucleic Acids Res.* **38**, D396-400.

Delcher, A.L., Bratke, K.A., Powers, E.C., and Salzberg, S.L. (2007). Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* **23**, 673–679.

Deng, S., Stein, R.A., and Higgins, N.P. (2005). Organization of supercoil domains and their reorganization by transcription. *Mol. Microbiol.* **57**, 1511–1521.

Denœud, F., and Vergnaud, G. (2004). Identification of polymorphic tandem repeats by direct comparison of genome sequence from different bacterial strains : a web-based resource. *BMC Bioinformatics* **5**, 4.

Despalins, A., Marsit, S., and Oberto, J. (2011). Absynte: a web tool to analyze the evolution of orthologous archaeal and bacterial gene clusters. *Bioinforma. Oxf. Engl.* **27**, 2905–2906.

D'haeseleer, P. (2006). What are DNA sequence motifs? *Nat. Biotechnol.* **24**, 423–425.

Dillon, S.C., and Dorman, C.J. (2010). Bacterial nucleoid-associated proteins, nucleoid structure and gene expression. *Nat. Rev. Microbiol.* **8**, 185–195.

Dinger, M.E., Gascoigne, D.K., and Mattick, J.S. (2011). The evolution of RNAs with multiple functions. *Biochimie* **93**, 2013–2018.

Dittmar, K.A., Sørensen, M.A., Elf, J., Ehrenberg, M., and Pan, T. (2005). Selective charging of tRNA isoacceptors induced by amino-acid starvation. *EMBO Rep.* **6**, 151–157.

El Karoui, M., Biaudet, V., Schbath, S., and Gruss, A. (1999). Characteristics of Chi distribution on different bacterial genomes. *Res. Microbiol.* **150**, 579–587.

Elf, J., Nilsson, D., Tenson, T., and Ehrenberg, M. (2003). Selective charging of tRNA isoacceptors explains patterns of codon usage. *Science* **300**, 1718–1722.

Erill, I., and O'Neill, M.C. (2009). A reexamination of information theory-based methods for DNA-binding site identification. *BMC Bioinformatics* **10**, 57.

Feklístov, A., Sharon, B.D., Darst, S.A., and Gross, C.A. (2014). Bacterial sigma factors: a historical, structural, and genomic perspective. *Annu. Rev. Microbiol.* **68**, 357–376.

Fong, C., Rohmer, L., Radey, M., Wasnick, M., and Brittnacher, M. (2008). PSAT: A web tool to compare genomic neighborhoods of multiple prokaryotic genomes. *BMC Bioinformatics* **9**, 170.

Francke, C., Kerkhoven, R., Wels, M., and Siezen, R.J. (2008). A generic approach to identify Transcription Factor-specific operator motifs; Inferences for LacI-family mediated regulation in *Lactobacillus plantarum* WCFS1. *BMC Genomics* **9**, 145.

Francke, C., Groot Kormelink, T., Hagemeijer, Y., Overmars, L., Sluijter, V., Moezelaar, R., and Siezen, R.J. (2011). Comparative analyses imply that the enigmatic sigma factor 54 is a central controller of the bacterial exterior. *BMC Genomics* **12**, 385.

Frith, M.C., Fu, Y., Yu, L., Chen, J.-F., Hansen, U., and Weng, Z. (2004). Detection of functional DNA motifs via statistical over-representation. *Nucleic Acids Res.* **32**, 1372–1381.

Frith, M.C., Saunders, N.F.W., Kobe, B., and Bailey, T.L. (2008). Discovering Sequence Motifs with Arbitrary Insertions and Deletions. *PLoS Comput Biol* **4**, e1000071.

Fujita, P.A., Rhead, B., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Cline, M.S., Goldman, M., Barber, G.P., Clawson, H., Coelho, A., et al. (2011). The UCSC Genome Browser database: update 2011. *Nucleic Acids Res.* **39**, D876-882.

Grin, I., and Linke, D. (2011). GCView: the genomic context viewer for protein homology searches. *Nucleic Acids Res*. **39**, W353-356.

Grissa, I., Vergnaud, G., and Pourcel, C. (2007a). The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics* **8**, 172.

Grissa, I., Vergnaud, G., and Pourcel, C. (2007b). CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res.* **35**, W52–W57.

Gruber, T.M., and Gross, C.A. (2003). Multiple sigma subunits and the partitioning of bacterial transcription space. *Annu. Rev. Microbiol.* **57**, 441–466.

Higgins, C.F., Ames, G.F.-L., Barnes, W.M., Clement, J.M., and Hofnung, M. (1982). A novel intercistronic regulatory element of prokaryotic operons. *Nature* **298**, 760–762.

Hu, G.-Q., Zheng, X., Yang, Y.-F., Ortet, P., She, Z.-S., and Zhu, H. (2008). ProTISA: a comprehensive resource for translation initiation site annotation in prokaryotic genomes. *Nucleic Acids Res.* **36**, D114–D119.

Hyatt, D., Chen, G.-L., LoCascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119.

Jansen, R., van Embden, J.D.A., Gaastra, W., and Schouls, L.M. (2002). Identification of a novel family of sequence repeats among prokaryotes. *Omics J. Integr. Biol.* **6**, 23–33.

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K.R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., et al. (2013). DNA-binding specificities of human transcription factors. *Cell* **152**, 327–339.

Jong, A. de, Pietersma, H., Cordes, M., Kuipers, O.P., and Kok, J. (2012). PePPER: a webserver for prediction of prokaryote promoter elements and regulons. *BMC Genomics* **13**, 299.

Kingsford, C.L., Ayanbule, K., and Salzberg, S.L. (2007). Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. *Genome Biol*. **8**, R22.

Kleckner, N., Fisher, J.K., Stouf, M., White, M.A., Bates, D., and Witz, G. (2014). The bacterial nucleoid: nature, dynamics and sister segregation. *Curr. Opin. Microbiol.* **22**, 127–137.

Koonin, E.V., and Wolf, Y.I. (2008). Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Res.* **36**, 6688–6719.

Kuhlman, T.E., and Cox, E.C. (2012). Gene location and DNA density determine transcription factor distributions in *Escherichia coli. Mol. Syst. Biol.* **8**, 610.

Ladunga, I. (2010). An overview of the computational analyses and discovery of transcription factor binding sites. *Methods Mol. Biol.* **674**, 1–22.

Larsen, T.S., and Krogh, A. (2003). EasyGene – a prokaryotic gene finder that ranks ORFs by statistical significance. *BMC Bioinformatics* **4**, 21.

Liu, M., Prakash, C., Nauta, A., Siezen, R.J., and Francke, C. (2012). A computational analysis of cysteine and methionine metabolism and its regulation in dairy starter and related bacteria. *J. Bacteriol.* JB.06816-11.

Ma, J., Campbell, A., and Karlin, S. (2002). Correlations between Shine-Dalgarno Sequences and Gene Features Such as Predicted Expression Levels and Operon Structures. *J. Bacteriol.* **184,** 5733–5745.

Mahillon, J., and Chandler, M. (1998). Insertion Sequences. Microbiol. *Mol. Biol. Rev.* **62**, 725–774.

Mandal, M., and Breaker, R.R. (2004). Gene regulation by riboswitches. *Nat. Rev. Mol. Cell Biol.* **5**, 451–463.

Marzi, S., Fechter, P., Chevalier, C., Romby, P., and Geissmann, T. (2008). RNA switches regulate initiation of translation in bacteria. *Biol. Chem.* **389**, 585–598.

Mathelier, A., Zhao, X., Zhang, A.W., Parcy, F., Worsley-Hunt, R., Arenillas, D.J., Buchman, S., Chen, C., Chou, A., Ienasescu, H., et al. (2013). JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.* gkt997.

McCue, L.A., Thompson, W., Carmack, C.S., Ryan, M.P., Liu, J.S., Derbyshire, V., and Lawrence, C.E. (2001). Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic Acids Res.* **29**, 774–782.

Menzel, R., and Gellert, M. (1983). Regulation of the genes for *E. coli* DNA gyrase: homeostatic control of DNA supercoiling. *Cell* **34**, 105–113.

Michael Gromiha, M., Siebers, J.G., Selvaraj, S., Kono, H., and Sarai, A. (2004). Intermolecular and intramolecular readout mechanisms in protein-DNA recognition. *J. Mol. Biol.* **337**, 285–294.

Mrázek, J. (2009). Finding sequence motifs in prokaryotic genomes—a brief practical guide for a microbiologist. *Brief. Bioinform.* bbp032.

Mrázek, J., Xie, S., Guo, X., and Srivastava, A. (2008). AIMIE: a web-based environment for detection and interpretation of significant sequence motifs in prokaryotic genomes. *Bioinformatics* **24**, 1041–1048.

Muto, A., and Osawa, S. (1987). The guanine and cytosine content of genomic DNA and bacterial evolution. *Proc. Natl. Acad. Sci. U. S. A.* **84**, 166–169.

Naville, M., and Gautheret, D. (2010a). Transcription attenuation in bacteria: theme and variations. *Brief. Funct. Genomics* **9**, 178–189.

Naville, M., and Gautheret, D. (2010b). Premature terminator analysis sheds light on a hidden world of bacterial transcriptional attenuation. *Genome Biol.* **11**, R97.

Nöllmann, M., Crisona, N.J., and Arimondo, P.B. (2007). Thirty years of *Escherichia coli* DNA gyrase: From in vivo function to single-molecule mechanism. *Biochimie* **89**, 490–499.

Nordberg, H., Cantor, M., Dusheyko, S., Hua, S., Poliakov, A., Shabalov, I., Smirnova, T., Grigoriev, I.V., and Dubchak, I. (2014). The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. *Nucleic Acids Res.* **42**, D26–D31.

Novichkov, P.S., Kazakov, A.E., Ravcheev, D.A., Leyn, S.A., Kovaleva, G.Y., Sutormin, R.A., Kazanov, M.D., Riehl, W., Arkin, A.P., Dubchak, I., et al. (2013). RegPrecise 3.0 – A resource for genome-scale exploration of transcriptional regulation in bacteria. *BMC Genomics* **14**, 745.

Overmars, L., Kerkhoven, R., Siezen, R.J., and Francke, C. (2013). MGcV: the microbial genomic context viewer for comparative genome analysis. *BMC Genomics* **14**, 209.

Overmars, L., van Hijum, S.A.F.T., Siezen, R.J., and Francke, C. (2015). CiVi: circular genome visualization with unique features to analyze sequence elements. *Bioinforma. Oxf. Engl.* **31**, 2867–2869.

Paget, M.S. (2015). Bacterial Sigma Factors and Anti-Sigma Factors: Structure, Function and Distribution. *Biomolecules* **5**, 1245–1265.

Paget, M.S.B., and Helmann, J.D. (2003). The sigma70 family of sigma factors. *Genome Biol.* **4**, 203.

Papenfort, K., and Vogel, J. (2010). Regulatory RNA in bacterial pathogens. *Cell Host Microbe* **8**, 116–127.

Pati, A., Ivanova, N.N., Mikhailova, N., Ovchinnikova, G., Hooper, S.D., Lykidis, A., and Kyrpides, N.C. (2010). GenePRIMP: a gene prediction improvement pipeline for prokaryotic genomes. *Nat. Methods* **7**, 455–457.

Postow, L., Hardy, C.D., Arsuaga, J., and Cozzarelli, N.R. (2004). Topological domain structure of the *Escherichia coli* chromosome. *Genes Dev.* **18**, 1766–1779.

Price, A.L., Jones, N.C., and Pevzner, P.A. (2005). *De novo* identification of repeat families in large genomes. *Bioinformatics* **21**, i351–i358.

Robinow, C., and Kellenberger, E. (1994). The bacterial nucleoid revisited. *Microbiol. Rev.* **58**, 211–232.

Rosenbloom, K.R., Armstrong, J., Barber, G.P., Casper, J., Clawson, H., Diekhans, M., Dreszer, T.R., Fujita, P.A., Guruvadoo, L., Haeussler, M., et al. (2015). The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res.* **43**, D670–D681.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.A., and Barrell, B. (2000). Artemis: sequence visualization and annotation. *Bioinforma. Oxf. Engl.* **16**, 944–945.

Sadler, J.R., Sasmor, H., and Betz, J.L. (1983). A perfectly symmetric lac operator binds the lac repressor very tightly. *Proc. Natl. Acad. Sci. U. S. A.* **80**, 6785–6789.

Serganov, A., and Nudler, E. (2013). A Decade of Riboswitches. *Cell* **152**, 17–24.

Siddharthan, R., Siggia, E.D., and van Nimwegen, E. (2005). PhyloGibbs: A Gibbs Sampling Motif Finder That Incorporates Phylogeny. *PLoS Comput Biol* **1**, e67.

Sierro, N., Makita, Y., Hoon, M. de, and Nakai, K. (2008). DBTBS: a database of transcriptional regulation in *Bacillus subtilis* containing upstream intergenic conservation information. *Nucleic Acids Res.* **36**, D93–D96.

Siguier, P., Perochon, J., Lestrade, L., Mahillon, J., and Chandler, M. (2006). ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res.* **34**, D32-36.

Siguier, P., Gourbeyre, E., and Chandler, M. (2014). Bacterial insertion sequences: their genomic impact and diversity. *FEMS Microbiol. Rev.* **38**, 865–891.

Siguier, P., Gourbeyre, E., Varani, A., Ton-Hoang, B., and Chandler, M. (2015). Everyman's Guide to Bacterial Insertion Sequences. *Microbiol. Spectr.* **3**, MDNA3-0030-2014.

Smith, G.R. (2012). How RecBCD enzyme and Chi promote DNA break repair and recombination: a molecular biologist's view. *Microbiol. Mol. Biol. Rev.* MMBR 76, 217–228.

Srividhya, K.V., Alaguraj, V., Poornima, G., Kumar, D., Singh, G.P., Raghavenderan, L., Katta, A.V.S.K.M., Mehta, P., and Krishnaswamy, S. (2007). Identification of Prophages in Bacterial Genomes by Dinucleotide Relative Abundance Difference. *PLoS ONE* **2**, e1193.

Stein, R.A., Deng, S., and Higgins, N.P. (2005). Measuring chromosome dynamics on different time scales using resolvases with varying half-lives. *Mol. Microbiol.* **56**, 1049–1061.

Stock, A.M., Robinson, V.L., and Goudreau, P.N. (2000). Two-component signal transduction. *Annu. Rev. Biochem.* **69**, 183–215.

Tech, M., Pfeifer, N., Morgenstern, B., and Meinicke, P. (2005). TICO: a tool for improving predictions of prokaryotic translation initiation sites. *Bioinformatics* **21**, 3568–3569.

Thomas-Chollier, M., Defrance, M., Medina-Rivera, A., Sand, O., Herrmann, C., Thieffry, D., and Helden, J. van (2011). RSAT 2011: regulatory sequence analysis tools. *Nucleic Acids Res.* **39**, W86–W91.

Tobes, R., and Ramos, J.-L. (2005). REP code: defining bacterial identity in extragenic space. *Environ. Microbiol.* **7**, 225–228.

Touzain, F., Petit, M.-A., Schbath, S., and Karoui, M.E. (2011). DNA motifs that sculpt the bacterial chromosome. *Nat. Rev. Microbiol.* **9**, 15–26.

Treangen, T.J., Abraham, A.-L., Touchon, M., and Rocha, E.P.C. (2009a). Genesis, effects and fates of repeats in prokaryotic genomes. *FEMS Microbiol. Rev.* **33**, 539–571.

Treangen, T.J., Darling, A.E., Achaz, G., Ragan, M.A., Messeguer, X., and Rocha, E.P.C. (2009b). A Novel Heuristic for Local Multiple Alignment of Interspersed DNA Repeats. *IEEEACM Trans Comput Biol Bioinforma.* **6**, 180–189.

Ulrich, L.E., Koonin, E.V., and Zhulin, I.B. (2005). One-component systems dominate signal transduction in prokaryotes. *Trends Microbiol.* **13**, 52–56.

Vallenet, D., Belda, E., Calteau, A., Cruveiller, S., Engelen, S., Lajus, A., Fèvre, F.L., Longin, C., Mornico, D., Roche, D., et al. (2013). MicroScope—an integrated microbial resource for the curation and comparative analysis of genomic and metabolic data. *Nucleic Acids Res.* **41**, D636–D647.

Vitreschak, A.G., Rodionov, D.A., Mironov, A.A., and Gelfand, M.S. (2003). Regulation of the vitamin B12 metabolism and transport in bacteria by a conserved RNA structural element. *RNA* **9**, 1084–1097.

Vitreschak, A.G., Mironov, A.A., Lyubetsky, V.A., and Gelfand, M.S. (2008). Comparative genomic analysis of T-box regulatory systems in bacteria. *RNA* **14**, 717–735.

Volfovsky, N., Haas, B.J., and Salzberg, S.L. (2001). A clustering method for repeat analysis in DNA sequences. *Genome Biol.* **2**, RESEARCH0027.

Weinberg, Z., Wang, J.X., Bogue, J., Yang, J., Corbino, K., Moy, R.H., and Breaker, R.R. (2010). Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes. *Genome Biol.* **11**, R31.

Wels, M., Francke, C., Kerkhoven, R., Kleerebezem, M., and Siezen, R.J. (2006). Predicting cis-acting elements of *Lactobacillus plantarum* by comparative genomics with different taxonomic subgroups. *Nucleic Acids Res.* **34**, 1947–1958.

Wels, M., Kormelink, T.G., Kleerebezem, M., Siezen, R.J., and Francke, C. (2008). An *in silico* analysis of T-box regulated genes and T-box evolution in prokaryotes, with emphasis on prediction of substrate specificity of transporters. B*MC Genomics* **9**, 330.

Wels, M., Overmars, L., Francke, C., Kleerebezem, M., and Siezen, R.J. (2011). Reconstruction of the regulatory network of *Lactobacillus plantarum* WCFS1 on basis of correlated gene expression and conserved regulatory motifs. *Microb. Biotechnol.* **4**, 333–344.

Winkler, W., Nahvi, A., and Breaker, R.R. (2002). Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. *Nature* **419**, 952–956.

Zechiedrich, E.L., Khodursky, A.B., Bachellier, S., Schneider, R., Chen, D., Lilley, D.M., and Cozzarelli, N.R. (2000). Roles of topoisomerases in maintaining steady-state DNA supercoiling in *Escherichia coli. J. Biol. Chem.* **275**, 8103–8113.

Chapter 2a

MGcV: the microbial genomic context viewer for comparative genome analysis

Lex Overmars, Robert Kerkhoven, Roland J. Siezen, Christof Francke

BMC genomics, 2013, 14:209

Abstract

Background

Conserved gene context is used in many types of comparative genome analyses. It is used to provide leads on gene function, to guide the discovery of regulatory sequences, but also to aid in the reconstruction of metabolic networks. We present the Microbial Genomic context Viewer (MGcV), an interactive, web-based application tailored to strengthen the practice of manual comparative genome context analysis for bacteria.

Results

MGcV is a versatile, easy-to-use tool that renders a visualization of the genomic context of any set of selected genes, genes within a phylogenetic tree, genomic segments, or regulatory elements. It is tailored to facilitate laborious tasks such as the interactive annotation of gene function, the discovery of regulatory elements, or the sequence-based reconstruction of gene regulatory networks. We illustrate that MGcV can be used in gene function annotation by visually integrating information on prokaryotic genes, like their annotation as available from NCBI with other annotation data such as Pfam domains, sub-cellular location predictions and gene-sequence characteristics such as GC content. We also illustrate the usefulness of the interactive features that allow the graphical selection of genes to facilitate data gathering (e.g. upstream regions, ID's or annotation), in the analysis and reconstruction of transcription regulation. Moreover, putative regulatory elements and their corresponding scores or data from RNA-seq and microarray experiments can be uploaded, visualized and interpreted in (ranked-) comparative context maps. The ranked maps allow the interpretation of predicted regulatory elements and experimental data in light of each other.

Conclusion

MGcV advances the manual comparative analysis of genes and regulatory elements by providing fast and flexible integration of gene related data combined with straightforward data retrieval. MGcV is available at: http://mgcv.cmbi.ru.nl.

Background

The number of sequenced prokaryotic genomes keeps expanding at a rapid pace. As a result, much of the function annotation of genes and other sequence elements relies increasingly on automated pipelines. Despite this tendency, human interference remains indispensable to translate genomic data correctly to biological meaning. Gene context and its evolutionary conservation is one of the genomic properties that can greatly aid the related (manual) genome analyses. The gene context provides many clues concerning function and biological role of a gene in a prokaryote (Huynen et al., 2000; Wolf et al., 2001). Gene context data thus benefits the reconstruction of the metabolic network (Francke et al., 2005; Park et al., 2010; Thiele and Palsson, 2010). Moreover, conserved gene context can also be applied to guide the identification of regulatory elements and therewith the reconstruction of the transcription regulatory network (e.g. (Francke et al., 2008; Neph and Tompa, 2006; Rodionov, 2007; Zhang and Gerstein, 2003)).

From a practical point of view, a comprehensive visualization of genomics data and information on function facilitates the process of data integration, and thereby reduces the time needed for interpretation. There are several ways to achieve this goal, as reflected by the variety in genome browsers and annotation platforms that have been developed. Conventional genome browsers include for instance UCSC genome browser (Fujita et al., 2011), Artemis (Rutherford et al., 2000) and GBrowse (Podicheti et al., 2009). This type of genome browser is characterized by a generic, highly configurable setup (i.e. typically, users can upload their genomes in genbank- and/or gff3-format) and display genomic data in separate 'tracks'. On the other hand, resources such as IMG (Markowitz et al., 2012), Microscope (Vallenet et al., 2009), MicrobesOnline (Dehal et al., 2010) and the SEED (Overbeek et al., 2005) serve as annotation platforms by providing the user genomic data, analysis tools and visualization options. In 2004 we introduced the Microbial Genome Viewer (Kerkhoven et al., 2004). This web-based genome viewer allowed users to explore bacterial genomes in linear maps and create a genome-wide visualization of data in circular maps. Yet, other tools have a more specific focus. For instance, BAGET allows users to retrieve the gene-context for a single gene (Oberto, 2008), whereas GeConT 2 allows users to visualize the genomic context of query genes (Martinez-Guerrero et al., 2008). Some tools specifically address conservation of gene order between orthologous genes, also denoted as 'synteny'. For instance, GeneclusterViz (Pejaver et al., 2012), GCView (Grin and Linke, 2011), PSAT (Fong et al., 2008) and Absynte (Despalins et al., 2011) provide a local gene context comparison based on blast (−like) similarity searches.

In the public domain, various resources provide organism specific reconstructions of particular regulons through the integration of genome sequence data and stored motifs. Examples of these are PEPPER (de Jong et al., 2012), RegulonDB (Gama-Castro et al., 2011), RegTransBase (Kazakov et al., 2007), PRODORIC (Grote et al., 2009), RegPrecise (Novichkov et al., 2010), ProdoNet (Klein et al., 2008), FITBAR (Oberto, 2010), RegAnalyst (Sharma et al., 2009) and MicrobesOnline (Dehal et al., 2010). Most of these resources enable automated predictions of regulatory sites based on stored motifs collected from literature. Some resources also in addition allow for *de novo* motif discovery, using tools such as MEME (Bailey et al., 2009), Tmod (Sun et al., 2010) and GIMSAN (Ng and Keich, 2008), which were developed to identify significantly overrepresented sequence motifs.

The versatility of the above resources comes at the cost of some flexibility and speed. We have therefore developed the web-application MGcV, which aims specifically to serve as an integrative visual interface to speed up a manual genome analysis. MGcV is a light-weight and flexible viewer that provides: i) a comparative view of the genomic context for query genome segments, like genes, sets of genes, or (user defined-) gene trees; ii) the integration of information on gene function enriched with additional annotation data such Pfam domains and sub-cellular location-predictions within a single 'track'; iii) the possibility to visually select genes and extract diverse gene-linked information, like upstream regions, protein sequence or function annotation; and iv) the possibility to upload and integrate experimental data and user-defined regulatory elements in adaptable views. MGcV thus enables the exploitation of gene context information in the annotation of gene function, the analysis of the evolutionary conservation of that context, the recovery of associated regulatory elements and the ranked comparative view of the identified elements in combination with microarray- or RNA-seq data. Hereby MGcV provides a visual heart to the manual sequence-based analysis of gene-function and gene-regulation in bacteria.

Methods

Data resources

The genome and protein sequences, the associated gene identifiers and function annotations (e.g. trivial names, COG categories, protein names) of all publicly available bacterial genomes are obtained from the FTP server of NCBI RefSeq (ftp://ftp.uniprot.org/) (Pruitt et al., 2012). Uniprot accessions mapped to NCBI GI-codes are retrieved from the Uniprot FTP server (ftp://ftp.uniprot.org/) (Magrane and UniProt Consortium, 2011). Pfam domains are obtained from the FTP server of EBI (ftp://ftp.ebi.ac.uk/pub/databases/

Pfam/) (Punta et al., 2012). Gene-sequence characteristics like GC-content are calculated using in-house scripts. Sub-cellular location predictions are obtained from the PSORTdb website (http://db.psort.org/) (Yu et al., 2011). The data is updated on a weekly basis and stored in a local MySQL database to enable fast access. The microarray data that are used to illustrate the capabilities of MGcV in the second case study were taken from (Chen et al., 2010).

Implementation

MGcV is a web-application developed using a combination of python, javascript and SVG (Scalable Vector Graphics). We implemented MGcV as a single page application; the front-end makes server side calls through Jquery and AJAX and receives the response from the server. The interface consists of four boxes (see Fig. 1). From left to right, these include: i) data input; ii) map settings; iii) data import and; iv) data export. The user can provide four types of input, NCBI RefSeq GI's, NCBI locus tags, genomic positions (NCBI Refseq genome accession tab start tab stop), or a Newick-tree (the leaf labels must contain NCBI-GI code). In addition, the user can search for specific genes by providing for instance gene product or gene names or by performing a BLAST search. Query data supplied by the user is parsed and mapped to the corresponding gene context data using python scripts. Uploaded (phylogenetic) trees are processed with Newick utilities (Junier and Zdobnov, 2010), which is also used to create a visual representation of the tree in SVG format. For the COG (NCBI) and Pfam annotations in the gene context maps a color scheme was assigned by designating a unique color to each identifier. In a similar way colors were assigned to the different protein location predictions. The applied color schemes can be found in the legend. Gene-associated quantitative data (e.g. GC-content) are converted in a red-to-green gradient which is projected on top of the genes. Likewise, gene-associated quantitative data uploaded by the user (e.g. microarray- or RNAseq- data) are converted in a red-to-green gradient. These data are then projected in a horizontal bar below the genes to allow the visual integration with annotation data and regulatory element predictions. Generated maps can be downloaded in SVG, PNG or PDF format. The conversion of SVG to PNG and PDF is done using 'Batik Rasterizer' (http://xmlgraphics.apache.org/batik/tools/rasterizer.html). The interface and interactive maps allow the user to interact with the data. Map interactivity is achieved by ECMAscript; linked information on genes and other sequence elements can be inspected by mouse-over, whereas a mouse-click can be used to select genes for subsequent analysis and data retrieval. MGcV is operable in modern browsers like Firefox, Chrome and Internet Explorer, where for all browsers the latest version is recommended.

2a

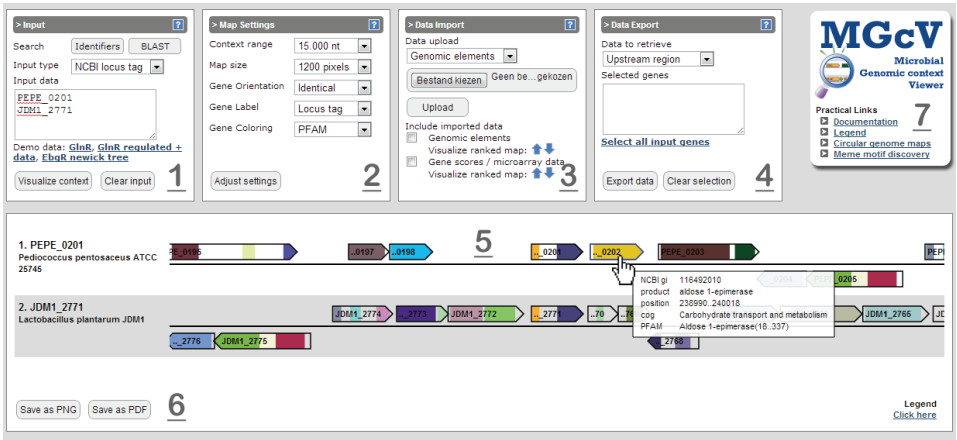


Figure 1. The interface of MGcV. The interface comprises seven separate modules. 1) Input: Copy/paste the input containing regions of interest and select the appropriate input-type. The context is visualized by clicking 'Visualize context'. Input-types include lists of identifiers such NCBI GI-codes or NCBI locus tags, genomic regions, or phylogenetic trees in newick format. Identifiers can also be searched for using the buttons 'Identifier' and 'BLAST'. 2) In the map settings, the appearance of the comparative context map can be altered. Various settings can be changed, which include the scale, the orientation, the gene-label and the data on display (e.g. COG coloring, PFAM, protein location). To apply the changes the 'Adjust settings'-button has to be clicked. 3) Data can be added via data import. Users can upload the position of genomic elements (e.g. regulatory elements) and gene-associated quantitative data (e.g. microarray data). Subsequently, the user can include the data in the map or generate a (ranked) comparative context map based on the uploaded data (blue arrows). 4) Sequence and functional data like upstream regions, gene sequences, cog categories, annotations can be obtained for selected genes via data export. Genes can be included or excluded by clicking. 5) The functional data related to a particular gene context is displayed in a single lane in the comparative genome map, thereby allowing for a direct comparison of the data between various strains or species. The maps are interactive; hovering the mouse over the map will display additional information (e.g. NCBI GI, gene product, COG) and clicking genes will tag them for data export. 6) The map can be converted to PNG- or PDF- format. 7) The interface is linked to a tutorial, to a color legend, to the circular viewer of the original MGV and also directly to the MEME website.

Results

Interface and functionality

Function annotation

The appropriate annotation of encoded function is essential for the correct interpretation of genomics data. The annotation process is initiated by the selection of genes and/or regions of interest. The flexible set-up of MGcV allows to generate an initial comparative context map simply by uploading a single identifier or a list of identifiers, like derived from a BLAST search, suffices to generate an initial comparative context map in MGcV. The uploaded identifiers may include NCBI gi-codes (RefSeq (Pruitt et al., 2012)), NCBI locus tags or genomic locations (designated by a RefSeq genome accession and position). In case the user does not have a list of gene identifiers, genes and their corresponding identifiers can be obtained via the built-in gene-search (input-box option 'Identifiers'). In addition, a BLAST search can be performed to find proteins similar to a given protein sequence. The BLAST hits can be selected and used as input for MGcV. We have also implemented the possibility to upload and visualize any (phylogenetic) gene tree. The combined view of gene phylogeny and the gene context allows a quick evaluation of the potential for similarity in molecular function and biological role between the selected genes. The labeling of the genes (i.e. by trivial name, by locus tag, or by NCBI GI-code), and similarly, the coloring of the genes (i.e. by COG category (Tatusov et al., 2003), by GC%, by sub cellular location (Yu et al., 2010) or by Pfam domain (Punta et al., 2012)) enhances the evaluation process. In addition, the genomic range of the maps can be altered and an identical orientation of the genes of interest can be enforced for purposes of presentation. The added value of MGcV in the manual function annotation is illustrated in more detail below (first case study).

Identification and comparison of regulatory elements

The starting point for a sequence-based reconstruction of transcription regulation is the identification of genes whose upstream region might contain a regulatory element, like a transcription factor (TF) binding site (e.g. (Neph and Tompa, 2006; Prakash and Tompa, 2005; Rodionov, 2007; Zhang and Gerstein, 2003)). We and others have shown that the identification of specific TF binding sites is particularly successful in the case of conserved gene context (e.g. (de Been et al., 2008; Francke et al., 2008; Rodionov and Gelfand, 2005)). We experienced that the ability to select upstream regions on basis of a visual representation of that context considerably speeds up the analysis and therefore have implemented this upstream region selection in MGcV.

Moreover, we have added a 'data import' option to allow the visualization of the predicted location of regulatory elements together with microarray or RNA-seq data. In this way, the location prediction of regulatory elements and the experimental data can be interpreted more easily in light of each other. In addition, the view can be ranked according to similarity score (for binding site predictions) or expression ratio (for microarray or RNA seq data). In fact, such a ranked view of expression data and gene context is also extremely useful in the interpretation of transcriptome experiments. The new features are illustrated below in the second case study.

Data export

An important aspect of data integration in comparative genome analyses is the combination of sequence and, sequence and function identifiers. Collecting these identifiers for a selected set of genes can be time-consuming; especially when the information linked to the genes found associated on the genome has to be included. We have added a 'data export' option in MGcV to accommodate the rapid and comprehensive collection of gene-related data. The user can graphically select genes of interest by mouse-click, where the selected genes are highlighted and included in the 'data export'-box. Subsequently, the data to be retrieved can be selected. These include for example upstream DNA sequences, protein sequences or function-related data like for instance: length, protein function, COG category or Pfam domains. The export option can be used without actually using the context view to, for instance, collect quickly the protein sequence or Uniprot accession codes for a set of gene IDs.

Case studies that illustrate the practical application of MGcV in manual comparative genome analysis

The main difference between MGcV and other resources is that MGvC is aimed to provide a platform to visually integrate one's own data (i.e. data generated externally using other tools or obtained through experimentation) with annotation data and practical export options that enable further (external) analysis. Other resources, like for instance MicrobesOnline (Dehal et al., 2010), in principle aim to offer a platform that is inclusive, i.e. that includes both calculation and visualization. Below we describe the results of two different manual comparative genome analysis using MGcV. In these two examples we highlight the flexible functionality of MGcV by visualizing the gene context and the associated functional information for a set of homologs that are present in a phylogenetic tree and by the visual integration of microarray data and *de novo* predictions of putative binding sites.

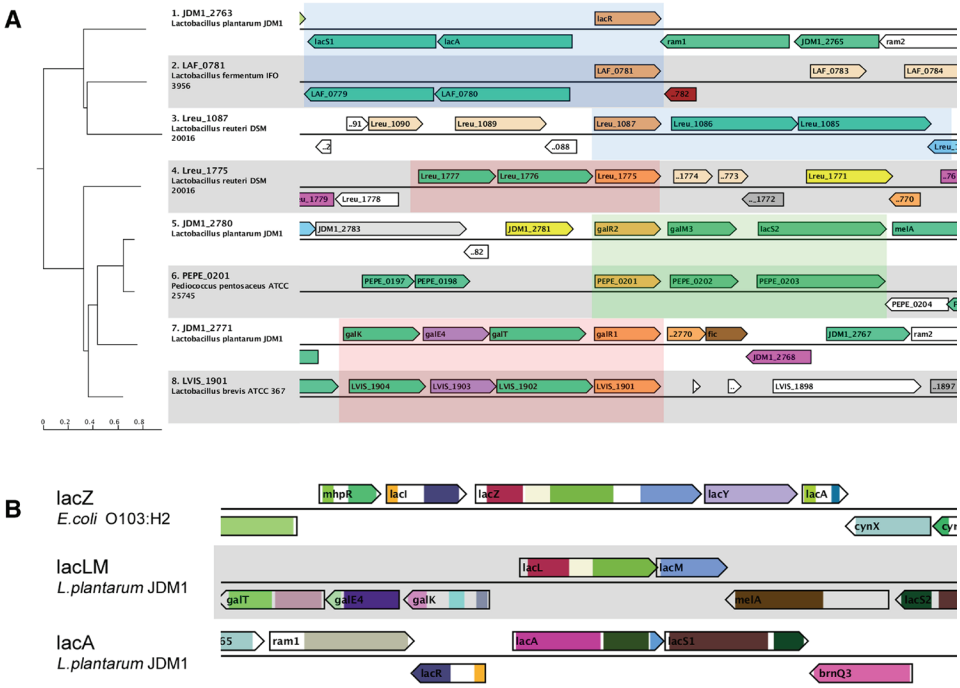
Case study 1: beta-galactosidases and associated regulators in Lactobacillus plantarum

The study of the *lac* operon and its control in *Escherichia coli* has set the paradigm of bacterial transcription regulation. The associated regulator in *E. coli* was named LacI. Most bacteria contain multiple homologs of this transcription factor family. In lactic acid bacteria, the *lac* operon is often associated with LacI-family regulators that form a separate clade within the family (e.g. EbgR in *E. coli*) (Francke et al., 2008). *L. plantarum* WCFS1 has three regulators that belong to this clade: LacR (ortholog in *B. subtilis* (Daniel et al., 1997)), GalR (ortholog in *S. thermophilus* (Ajdić and Ferretti, 1998)) and RafR (Silvestroni et al., 2002). To find functional equivalents in other *Lactobacilli*, the protein sequences of homologs were collected using a BLAST search. The recovered protein sequences were aligned and a neighbor-joining tree was constructed. To determine the degree of conservation the tree was used as input for MGcV. As shown in Figure 2A, integration with gene-context enhances both the interpretation of the tree and the identification of orthologs. Based on the integrated visualization of the phylogenetic tree and genomic context we can easily distinguish three different clusters. One of the first things that can be done on basis of the integrated view is a specification of annotation information as present in the NCBI database for orthologous genes that share context. The genes *JDM1_2771* (*galR1*) and *JDM_2780* (*galR2*) can be easily re-annotated to *galR* and *rafR*, respectively, on basis of the specific annotation that is available for *L. plantarum WCFS1*. Also the functional equivalency between genes can be evaluated, like for the gene *Lreu_1775*, which is the only regulator of e*bgR*-type in the *Lactobacillus reuterii* genome. Based on the tree and the fact that the gene has a similar gene context as *JDM1_2771* (*galR*) and *LVIS_1901* and not as *JDM_2780* (*rafR*) and *PEPE_0201,* it can be annotated as *galR* with the expected inducer galactose.

The production of galacto-oligosaccharides using microbial beta-galactosidases is currently well-studied in the field of functional foods (Park and Oh, 2010). In *Escherichia coli* a gene encoding beta-galactosidase: *lacZ*, was described first by Joshua Lederberg in 1948 (Lederberg, 1948). It took 25 years before a second beta-galactosidase encoding gene was described (Campbell et al., 1973), which was designated *ebgA* from evolved beta-galactosidase. The discovery resulted in the classic study (designation by (Dean, 2010)) of molecular evolution (review in (Hall, 2003)). The Pfam and COG classification (Figure 2B) comply with the assertion that both genes have evolved from a common ancestor. In many lactobacilli a third closely-related variant is found, *lacLM*. In some Lactobacilli (e.g. *L. delbrueckii* and *L. salivarius*) the protein is encoded by a single gene. However, in most

Lactobacilli the protein is encoded by two neighboring genes (probably the result of gene fission) and the active protein is a heterodimer (Nguyen et al., 2007a). It is the LacLM protein that is mostly exploited in biotechnological applications (Liu et al., 2011; Nguyen et al., 2007b). Like *E. coli*, various Lactobacilli have a second beta-galactosidase encoding gene, *lacA*. However, this gene has a completely different evolutionary origin and thus represents a functional analog. This conclusion can also easily be derived from the (pfam-) annotation information that is available in MGcV (Fig. 2B).

We have maintained the circular viewer of the original MGV in which we constructed a circular genome map of *L. plantarum* (Fig. 2C). In this map we included the locations of regulator-encoding genes *lacR*, *rafR* and *galR*, the GC-percentage and putative binding sites (similarity to motif >90% (Francke et al., 2008)). The genomic segment containing *lacR, rafR* and *galR* is flanking a region with a decreased GC-percentage, which was suggested to represent a lifestyle adaptation region in which many genes are acquired by horizontal gene transfer (Kleerebezem et al., 2003).



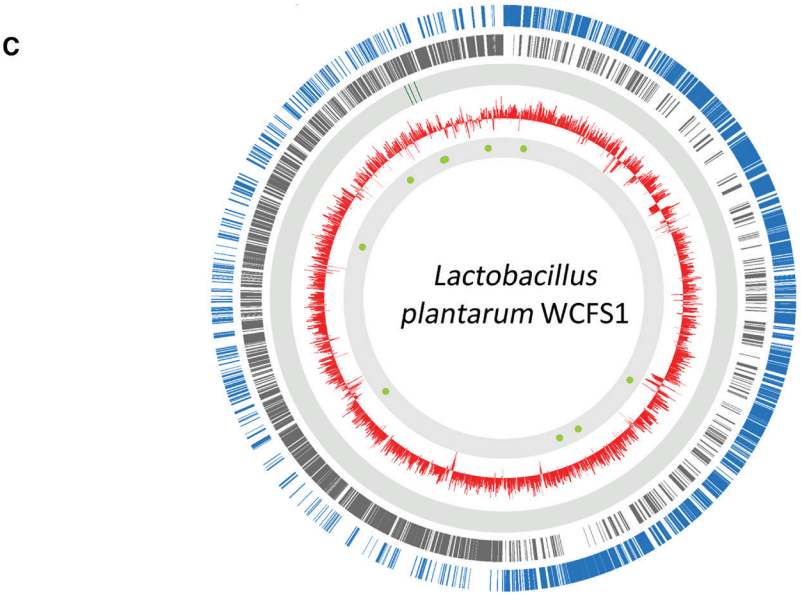


Figure 2. EbgR-like transcription factors in *L. plantarum* and other lactobacilli. A) MGcV visualization of a phylogenetic tree of EbgR-type regulators in some Lactobacilli. To simplify, the tree was pruned (species: *Lactobacillus plantarum JDM1* and *WCFS1*, *Lactobacillus fermentum IFO3956*, *Lactobacillus reuteri DSM20016*, *Lactobacillus brevis ATCC 367* and *Pediococcus pentosaceus*). The context range was set to 10.000 nucleotides, genes were colored by COG-class and trivial names were used to label genes. The visual combination of the phylogenetic tree and genomic context allows to distinguish three groups; lacR-, rafR- and galR-like sequences (designated in blue, green and red, respectively). **B**) Comparative context map of the beta-galactosidase encoding genes *lacZ* (*E. coli*), *lacLM* (*L. plantarum*) and *lacA* (*L. plantarum*). Pfam domains are used for gene-coloring and trivial names are used to label genes. In *L. plantarum lacLM* and *lacA* are both annotated to encode a protein with the same name (beta-galactosidase). Yet, from this comparative view it becomes clear that *lacLM* and *lacA* must have a different evolutionary origin. Although lacLM is encoded by two genes, the domain structure appears identical to the single *lacZ* gene of *E. coli*. **C**) A circular genome map of *L. plantarum* in which the ORFs on the plus strand (blue), on the minus stand (grey), the locations of regulator encoding genes *lacR*, *rafR* and *galR* (green), the GC% (red) and putative binding sites (similarity to motif >90% (Francke et al., 2008); represented by the green dots) are included.

Case study 2: Reconstruction of GlnR-mediated regulation in Streptococcus mutans

Recently, we have published a comparative genomics study on the transcription factor GlnR (Groot Kormelink et al., 2012). GlnR is one of the four major transcription factors involved in the control of central nitrogen metabolism in *Bacillus subtilis*. A BLAST search was performed to retrieve GlnR orthologs from all sequenced Streptococcal genomes and the gene context for the resulting list was displayed in MGcV (see Fig. 3A). We observed a clear conservation of the *glnRA* operon and its genomic context in all Streptococcaceae. MGcV was then used to collect selected upstream regions (Fig. 3B). These were analyzed using MEME (via the available link; (Bailey et al., 2009)) to search for a motif representing the GlnR-binding site. The motif was then refined and used to identify and score putative binding sites on all Streptococcal genomes (via e.g. MAST (Bailey et al., 2009) or a Similar Motif Search (Francke et al., 2011)). Subsequently, the resulting list of putative sites with their corresponding similarity scores was uploaded to MGcV. The view was ranked according to similarity score and the binding site predictions could be evaluated in light of their position relative to the genes. Then, the consistency of the predictions with microarray data was checked visually in MGcV. Chen and colleagues constructed a GlnR gene knockout in *Streptococcus mutans* for which they performed a microarray experiment (Chen et al., 2010). These data were retrieved and uploaded and the view was ranked according to microarray ratios (Fig. 3C). The view makes immediately clear that the predicted binding-sites are consistent with the microarray data. In addition, the view shows that the operon showing the strongest response (consisting of SMU_870, SMU_871 and SMU_872) is not preceded by a putative binding-site and therefore probably is regulated indirectly. In fact, this operon encodes a PTS system for which no functional relation with nitrogen is described. Interestingly, many of the high-scoring putative binding sites are followed by a binding site in the N-terminus encoding part of the gene (Fig. 3C: SMU_671 and SMU_1519), suggesting that this might be a particularity of the regulatory mechanism. Finally, the interactive map provides a convenient overview to determine a possible score threshold for both the predictions as well as the expression data.

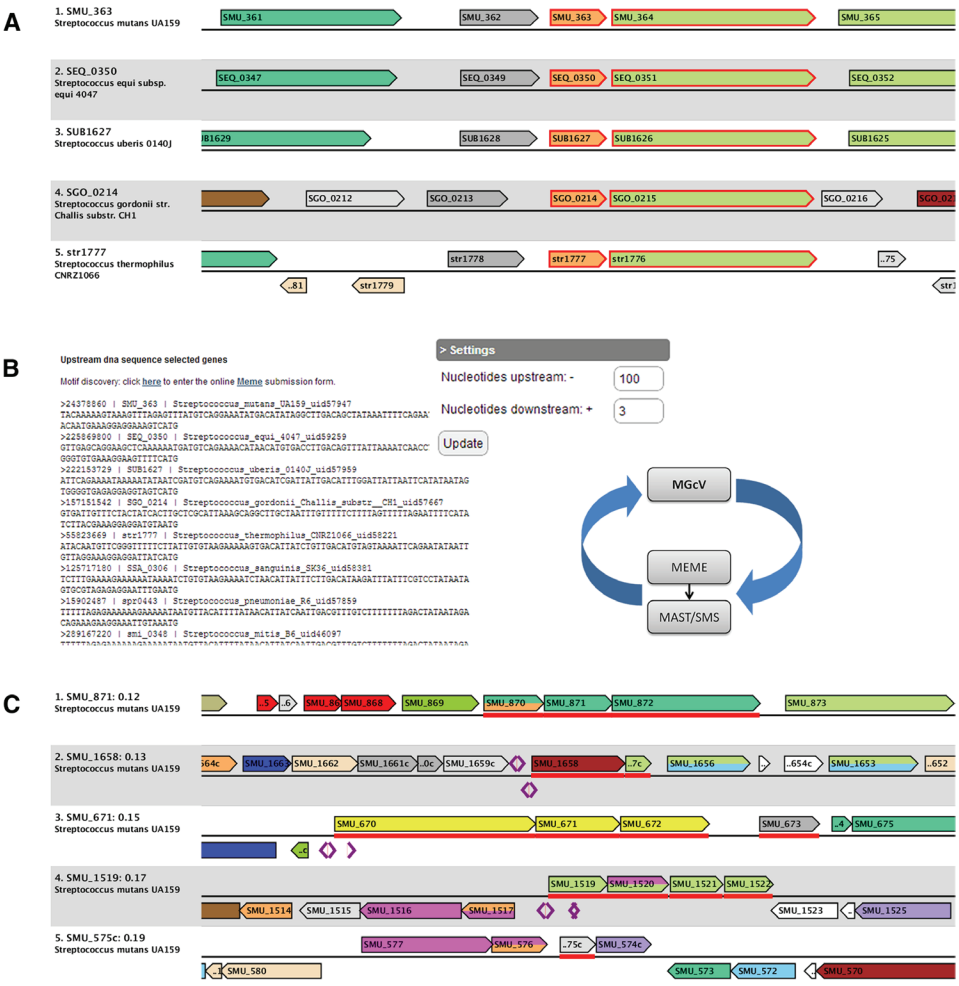


Figure 3. Application of MGcV in the reconstruction of GlnR-mediated regulation in *Streptococcus mutans*. **A**) Comparative context map of GlnR homologs obtained via a BLAST search in all sequenced *Streptococci* (only five species are shown). The GlnRA operon and its direct genomic context is clearly conserved in the Streptococci. The map was used to graphically select those genes that could be preceded by a binding site. **B**) The 'upstream region' option of the 'Data-export'-box was used to obtain upstream regions of the selected genes. Subsequently, the available link to MEME was used to search for possible overrepresented motifs. The results were refined and a motif defined (Francke et al., 2011), which was then used to search and score putative binding sites (e.g. using MAST or SMS). **C**) A comparative context map ranked on expression ratios (low-to-high) of a GlnR mutant, visualized in conjunction with predicted GlnR binding sites. To exemplify, the figure is limited to the top 5 of down-regulated genes. In this map, gene expression ratios are represented in a colored bar (red-to-green gradient; red is down-regulated) at the baseline and putative binding sites are designated by purple arrows (direction representing the strand). Both the microarray data and putative binding sites and their corresponding binding sites were uploaded using the 'Data import'-box. The resulting map allows the analysis of the putative GlnR binding sites in light of the expression data of the GlnR mutant. Most of the top down regulated genes (SMU_1658, SMU_671 and SMU_1519 in panel C) indeed are preceded by a putative binding site.

Conclusion

Gene-context conservation is an important genomic property to exploit in genome analyses. Nine years ago we developed a Microbial Genome Viewer (Kerkhoven et al., 2004) to support our efforts in the gene annotation and metabolic reconstruction of the lactic acid bacterium *Lactobacillus plantarum WCFS1* (Siezen et al., 2012; Teusink et al., 2005). Over the years we have experienced the need for additional functionality and more flexibility to enhance the work on the curation of function annotation and on the reconstruction of transcription regulatory networks. While maintaining the functionality, we have changed the complete setup and developed a new interface to create an adaptable interactive Microbial Genome context Viewer with high speed and versatile functionality to aid small-scale analyses. Both the input and output options of MGcV provide many practical features. The interactive maps allow users to graphically select sets of genes for data retrieval and subsequent analyses. Moreover, the maps provide a single integrated view of the data. The maps are made available in SVG, PNG and PDF format and are hereby suited to use as illustrations in publications, posters and presentations. The MGcV features that constitute its value to the manual analysis of genome sequence include: i) its light-weight and flexible interface; ii) the possibility to a) select multiple genes in the maps and extract gene-related data for these; and b) extract selected upstream regions to be used for further analysis; iii) the visual integration of a user-defined phylogenetic tree and the related gene context; and iv) the visual integration and ranking of microarray data or regulatory element predictions in the context of gene organization. Regarding the regulatory elements, any list of positions linked to a quantitative score can be uploaded, ranked and viewed. Possible applications of MGcV include: annotation refinement, function prediction on basis of a (phylogenetic) tree and conserved gene context, the sequence-based reconstruction of gene regulatory networks, and microarray/RNA-seq data analysis. We have presented two case studies to illustrate the practical applications of MGcV. Altogether, MGcV provides a flexible platform to exploit publicly available genomic data in small scale genome analysis in a fast and convenient manner.

Acknowledgements

This work was part of the BioRange programme of the Netherlands Bioinformatics Centre (NBIC), which is supported by the Netherlands Genomics Initiative (NGI). We thank Marieke Bart, Tom Groot Kormelink, Lennart Backus and Mark de Been for their contributions to the project.

References

Ajdić, D., and Ferretti, J.J. (1998). Transcriptional regulation of the *Streptococcus mutans* gal operon by the GalR repressor. *J. Bacteriol.* **180**, 5727–5732.

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202-208.

de Been, M., Bart, M.J., Abee, T., Siezen, R.J., and Francke, C. (2008). The identification of response regulator-specific binding sites reveals new roles of two-component systems in *Bacillus cereus* and closely related low-GC Gram-positives. *Environ. Microbiol.* **10**, 2796–2809.

Campbell, J.H., Lengyel, J.A., and Langridge, J. (1973). Evolution of a second gene for beta-galactosidase in *Escherichia coli. Proc. Natl. Acad. Sci. U. S. A*. **70**, 1841–1845.

Chen, P.-M., Chen, Y.-Y.M., Yu, S.-L., Sher, S., Lai, C.-H., and Chia, J.-S. (2010). Role of GlnR in acid-mediated repression of genes encoding proteins involved in glutamine and glutamate metabolism in *Streptococcus mutans. Appl. Environ. Microbiol.* **76**, 2478–2486.

Daniel, R.A., Haiech, J., Denizot, F., and Errington, J. (1997). Isolation and characterization of the lacA gene encoding beta-galactosidase in *Bacillus subtilis* and a regulator gene, lacR. *J. Bacteriol. 179*, 5636–5638.

Dean, A.M. (2010). The future of molecular evolution. *EMBO Rep*. **11**, 409.

Dehal, P.S., Joachimiak, M.P., Price, M.N., Bates, J.T., Baumohl, J.K., Chivian, D., Friedland, G.D., Huang, K.H., Keller, K., Novichkov, P.S., et al. (2010). MicrobesOnline: an integrated portal for comparative and functional genomics. *Nucleic Acids Res.* **38**, D396-400.

Despalins, A., Marsit, S., and Oberto, J. (2011). Absynte: a web tool to analyze the evolution of orthologous archaeal and bacterial gene clusters. *Bioinforma. Oxf. Engl.* **27**, 2905–2906.

Fong, C., Rohmer, L., Radey, M., Wasnick, M., and Brittnacher, M. (2008). PSAT: A web tool to compare genomic neighborhoods of multiple prokaryotic genomes. *BMC Bioinformatics* **9**, 170.

Francke, C., Siezen, R.J., and Teusink, B. (2005). Reconstructing the metabolic network of a bacterium from its genome. *Trends Microbiol.* **13**, 550–558.

Francke, C., Kerkhoven, R., Wels, M., and Siezen, R.J. (2008). A generic approach to identify Transcription Factor-specific operator motifs; Inferences for LacI-family mediated regulation in *Lactobacillus plantarum* WCFS1. *BMC Genomics* **9**, 145.

Francke, C., Groot Kormelink, T., Hagemeijer, Y., Overmars, L., Sluijter, V., Moezelaar, R., and Siezen, R.J. (2011). Comparative analyses imply that the enigmatic sigma factor 54 is a central controller of the bacterial exterior. *BMC Genomics* **12**, 385.

Fujita, P.A., Rhead, B., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Cline, M.S., Goldman, M., Barber, G.P., Clawson, H., Coelho, A., et al. (2011). The UCSC Genome Browser database: update 2011. *Nucleic Acids Res.* **39**, D876-882.

Gama-Castro, S., Salgado, H., Peralta-Gil, M., Santos-Zavaleta, A., Muñiz-Rascado, L., Solano-Lira, H., Jimenez-Jacinto, V., Weiss, V., García-Sotelo, J.S., López-Fuentes, A., et al. (2011). RegulonDB version 7.0: transcriptional regulation of *Escherichia coli* K-12 integrated within genetic sensory response units (Gensor Units). *Nucleic Acids Res.* **39**, D98-105.

Grin, I., and Linke, D. (2011). GCView: the genomic context viewer for protein homology searches. *Nucleic Acids Res*. **39**, W353-356.

Groot Kormelink, T., Koenders, E., Hagemeijer, Y., Overmars, L., Siezen, R.J., de Vos, W.M., and Francke, C. (2012). Comparative genome analysis of central nitrogen metabolism and its control by GlnR in the class *Bacilli*. *BMC Genomics* **13**, 191.

Grote, A., Klein, J., Retter, I., Haddad, I., Behling, S., Bunk, B., Biegler, I., Yarmolinetz, S., Jahn, D., and Münch, R. (2009). PRODORIC (release 2009): a database and tool platform for the analysis of gene regulation in prokaryotes. *Nucleic Acids Res.* **37**, D61-65.

Hall, B.G. (2003). The EBG system of *E. coli*: origin and evolution of a novel beta-galactosidase for the metabolism of lactose. *Genetica* **118**, 143–156.

Huynen, M., Snel, B., Lathe, W., and Bork, P. (2000). Exploitation of gene context. *Curr. Opin. Struct. Biol.* **10**, 366–370.

de Jong, A., Pietersma, H., Cordes, M., Kuipers, O.P., and Kok, J. (2012). PePPER: a webserver for prediction of prokaryote promoter elements and regulons. *BMC Genomics* **13**, 299.

Junier, T., and Zdobnov, E.M. (2010). The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. *Bioinforma. Oxf. Engl.* **26**, 1669–1670.

Kazakov, A.E., Cipriano, M.J., Novichkov, P.S., Minovitsky, S., Vinogradov, D.V., Arkin, A., Mironov, A.A., Gelfand, M.S., and Dubchak, I. (2007). RegTransBase--a database of regulatory sequences and interactions in a wide range of prokaryotic genomes. *Nucleic Acids Res.* **35,** D407-412.

Kerkhoven, R., van Enckevort, F.H.J., Boekhorst, J., Molenaar, D., and Siezen, R.J. (2004). Visualization for genomics: the Microbial Genome Viewer. *Bioinformatics* **20,** 1812–1814.

Kleerebezem, M., Boekhorst, J., van Kranenburg, R., Molenaar, D., Kuipers, O.P., Leer, R., Tarchini, R., Peters, S.A., Sandbrink, H.M., Fiers, M.W.E.J., et al. (2003). Complete genome sequence of *Lactobacillus plantarum* WCFS1. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 1990–1995.

Klein, J., Leupold, S., Münch, R., Pommerenke, C., Johl, T., Kärst, U., Jänsch, L., Jahn, D., and Retter, I. (2008). ProdoNet: identification and visualization of prokaryotic gene regulatory and metabolic networks. *Nucleic Acids Res.* **36**, W460-464.

Lederberg, J. (1948). Gene control of beta-galactosidase in *Escherichia coli*. *Genetics* **33**, 617.

Liu, G.X., Kong, J., Lu, W.W., Kong, W.T., Tian, H., Tian, X.Y., and Huo, G.C. (2011). β-Galactosidase with transgalactosylation activity from *Lactobacillus fermentum* K4. *J. Dairy Sci.* **94**, 5811–5820.

Magrane, M., and UniProt Consortium (2011). UniProt Knowledgebase: a hub of integrated protein data. *Database (Oxford)* **2011,** bar009.

Markowitz, V.M., Chen, I.-M.A., Palaniappan, K., Chu, K., Szeto, E., Grechkin, Y., Ratner, A., Jacob, B., Huang, J., Williams, P., et al. (2012). IMG: the Integrated Microbial Genomes database and comparative analysis system. *Nucleic Acids Res.* **40**, D115-122.

Martinez-Guerrero, C.E., Ciria, R., Abreu-Goodger, C., Moreno-Hagelsieb, G., and Merino, E. (2008). GeConT 2: gene context analysis for orthologous proteins, conserved domains and metabolic pathways. *Nucleic Acids Res.* **36**, W176-180.

Neph, S., and Tompa, M. (2006). MicroFootPrinter: a tool for phylogenetic footprinting in prokaryotic genomes. *Nucleic Acids Res.* **34**, W366-368.

Ng, P., and Keich, U. (2008). GIMSAN: a Gibbs motif finder with significance analysis. *Bioinforma. Oxf. Engl.* **24**, 2256–2257.

Nguyen, T.-H., Splechtna, B., Krasteva, S., Kneifel, W., Kulbe, K.D., Divne, C., and Haltrich, D. (2007a). Characterization and molecular cloning of a heterodimeric beta-galactosidase from the probiotic strain *Lactobacillus acidophilus* R22. *FEMS Microbiol. Lett.* **269**, 136–144.

Nguyen, T.-H., Splechtna, B., Yamabhai, M., Haltrich, D., and Peterbauer, C. (2007b). Cloning and expression of the beta-galactosidase genes from *Lactobacillus reuteri* in *Escherichia coli*. *J. Biotechnol.* **129**, 581–591.

Novichkov, P.S., Laikova, O.N., Novichkova, E.S., Gelfand, M.S., Arkin, A.P., Dubchak, I., and Rodionov, D.A. (2010). RegPrecise: a database of curated genomic inferences of transcriptional regulatory interactions in prokaryotes. *Nucleic Acids Res.* **38**, D111-118.

Oberto, J. (2008). BAGET: a web server for the effortless retrieval of prokaryotic gene context and sequence. *Bioinforma. Oxf. Engl.* **24**, 424–425.

Oberto, J. (2010). FITBAR: a web tool for the robust prediction of prokaryotic regulons. *BMC Bioinformatics* **11**, 554.

Overbeek, R., Begley, T., Butler, R.M., Choudhuri, J.V., Chuang, H.-Y., Cohoon, M., de Crécy-Lagard, V., Diaz, N., Disz, T., Edwards, R., et al. (2005). The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.* **33**, 5691–5702.

Park, A.-R., and Oh, D.-K. (2010). Galacto-oligosaccharide production using microbial beta-galactosidase: current state and perspectives. *Appl. Microbiol. Biotechnol.* **85**, 1279–1286.

Park, J.M., Kim, T.Y., and Lee, S.Y. (2010). Prediction of metabolic fluxes by incorporating genomic context and flux-converging pattern analyses. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 14931–14936.

Pejaver, V.R., An, J., Rhee, S., Bhan, A., Choi, J.-H., Liu, B., Lee, H., Brown, P.J., Kysela, D., Brun, Y.V., et al. (2012). GeneclusterViz: a tool for conserved gene cluster visualization, exploration and analysis. *Bioinforma. Oxf. Engl.* **28,** 1527–1529.

Podicheti, R., Gollapudi, R., and Dong, Q. (2009). WebGBrowse--a web server for GBrowse. *Bioinforma. Oxf. Engl.* **25**, 1550–1551.

Prakash, A., and Tompa, M. (2005). Discovery of regulatory elements in vertebrates through comparative genomics. *Nat. Biotechnol.* **23**, 1249–1256.

Pruitt, K.D., Tatusova, T., Brown, G.R., and Maglott, D.R. (2012). NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.* **40**, D130-135.

Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., et al. (2012). The Pfam protein families database. *Nucleic Acids Res.* **40**, D290-301.

Rodionov, D.A. (2007). Comparative genomic reconstruction of transcriptional regulatory networks in bacteria. *Chem. Rev.* **107**, 3467–3497.

Rodionov, D.A., and Gelfand, M.S. (2005). Identification of a bacterial regulatory system for ribonucleotide reductases by phylogenetic profiling. *Trends Genet.* TIG **21**, 385–389.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.A., and Barrell, B. (2000). Artemis: sequence visualization and annotation. *Bioinforma. Oxf. Engl.* **16**, 944–945.

Sharma, D., Mohanty, D., and Surolia, A. (2009). RegAnalyst: a web interface for the analysis of regulatory motifs, networks and pathways. *Nucleic Acids Res.* **37**, W193-201.

Siezen, R.J., Francke, C., Renckens, B., Boekhorst, J., Wels, M., Kleerebezem, M., and van Hijum, S.A.F.T. (2012). Complete resequencing and reannotation of the *Lactobacillus plantarum* WCFS1 genome. *J. Bacteriol.* **194**, 195–196.

Silvestroni, A., Connes, C., Sesma, F., De Giori, G.S., and Piard, J.-C. (2002). Characterization of the melA locus for alpha-galactosidase in *Lactobacillus plantarum*. *Appl. Environ. Microbiol.* **68**, 5464–5471.

Sun, H., Yuan, Y., Wu, Y., Liu, H., Liu, J.S., and Xie, H. (2010). Tmod: toolbox of motif discovery. *Bioinforma. Oxf. Engl.* **26**, 405–407.

Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., et al. (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41.

Teusink, B., van Enckevort, F.H.J., Francke, C., Wiersma, A., Wegkamp, A., Smid, E.J., and Siezen, R.J. (2005). In Silico Reconstruction of the Metabolic Pathways of *Lactobacillus plantarum*: Comparing Predictions of Nutrient Requirements with Those from Growth Experiments. *Appl Env. Microbiol* **71**, 7253–7262.

Thiele, I., and Palsson, B.Ø. (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat. Protoc.* **5**, 93–121.

Vallenet, D., Engelen, S., Mornico, D., Cruveiller, S., Fleury, L., Lajus, A., Rouy, Z., Roche, D., Salvignol, G., Scarpelli, C., et al. (2009). MicroScope: a platform for microbial genome annotation and comparative genomics. *Databases (Oxford)* **2009**, bap021.

Wolf, Y.I., Rogozin, I.B., Kondrashov, A.S., and Koonin, E.V. (2001). Genome Alignment, Evolution of Prokaryotic Genome Organization, and Prediction of Gene Function Using Genomic Context. *Genome Res.* **11**, 356–372.

Yu, N.Y., Wagner, J.R., Laird, M.R., Melli, G., Rey, S., Lo, R., Dao, P., Sahinalp, S.C., Ester, M., Foster, L.J., et al. (2010). PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics* **26**, 1608–1615.

Yu, N.Y., Laird, M.R., Spencer, C., and Brinkman, F.S.L. (2011). PSORTdb--an expanded, auto-updated, user-friendly protein subcellular localization database for Bacteria and Archaea. *Nucleic Acids Res.* **39**, D241-244.

Zhang, Z., and Gerstein, M. (2003). Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements. *J. Biol.* **2**, 11.

Chapter 2b

CiVi: circular genome visualization with unique features to analyze sequence elements

Lex Overmars, Sacha A.F.T. van Hijum, Roland J. Siezen, Christof Francke

Bioinformatics, 2015, 31: 2867-2869.

Abstract

Summary

We have developed CiVi, a user-friendly web-based tool to create custom circular maps to aid the analysis of microbial genomes and sequence elements. Sequence related data such as gene-name, COG class, PFAM domain, GC%, and subcellular location can be comprehensively viewed. Quantitative gene-related data (e.g. expression ratios or read counts) as well as predicted sequence elements (e.g. regulatory sequences) can be uploaded and visualized. CiVi accommodates the analysis of genomic elements by allowing a visual interpretation in the context of: i) their genome-wide distribution, ii) provided experimental data and iii) the local orientation and location with respect to neighboring genes. CiVi thus enables both experts and non-experts to conveniently integrate public genome data with the results of genome analyses in circular genome maps suitable for publication.

Availability

CiVi is freely available at http://civi.cmbi.ru.nl

Introduction

Circular genome representations provide an excellent way to comprehensively inspect genome-wide data. Various tools have been developed to generate these circular visualizations including CGView (Stothard and Wishart 2005), GenomeVx (Conant and Wolfe 2008), GeneWiz (Hallin et al. 2009) and DNAPlotter (Carver et al. 2009). Tools such as BRIG (Alikhan et al., 2011), CGView Comparison Tool (Grant et al., 2012), Circos (Krzywinski et al., 2009) and Circleator (Crabtree et al. 2014) can also visualize genome comparisons and in some cases, visualize links between genome sequence and other types of information. The tools described above are well-suited for the visualization of high-throughput genomic data like sequence-similarities or read counts, but have limited functionality in relation to smaller-scale activities such as reconstructing transcription networks and finding gene functions associated to genetic elements. Moreover, with the exception of GeneWiz (Hallin et al. 2009), they require laborious and sometimes complex uploads of genome and annotation data. In 2013 we published MGcV (Overmars et al., 2013), a linear- genomic context visualization tool tailored to provide a simple and quick visual access to the publicly available genomic data from NCBI (Pruitt et al., 2012). The tool incorporated the capabilities of the earlier MGV (Kerkhoven et al., 2004), and extended the visualization with data export options to advance the gene-specific analysis of microbial genomes. We now present CiVi, which has been developed using the same philosophy, to extend the circular viewing options provided by the original MGV. The extensions include the possibility to display annotation data like COG category, PFAM domain or subcellular location directly on the genome map and to reveal the position of selected annotations through a keyword search option. CiVi also enables the upload and visualization of custom data, such as the positions of genomic elements and the export of associated data. The latter include the function annotations and/or sequences of neighboring genes, as well as information on the distance distribution of the elements with respect to the genes. The resulting circular maps can be edited and the pictures exported in svg-, png- and pdf-format. Finally, CiVi offers a completely new interface and back-end with enhanced usability, interactivity (via mouse-over) and speed.

Usage and implementation

CiVi enables users to create custom circular microbial genome maps in a simple step-wise fashion, adding data ring by ring. The interface consists of a panel on the right in which the map is displayed and four panels on the left related to the different menus, labeled: i) 'Genome and data selection'; ii) 'On display'; iii) 'Data import'; and iv) 'Elements and genomic context'. For every ring the user can subsequently set (panel i): the organism and genome of interest, the type of information, the coloring and the radius of the ring. The types of information that can be included directly are: the location of the genes on the +strand and -strand, COG categories (NCBI RefSeq), GC%, GC-skew and AT-skew (calculated; (Overmars et al., 2013)), PFAM domains, and subcellular location predictions (PSORTdb; (Nancy et al., 2010)). A keyword matching option allows highlighting genes whose gene product, gene name, COG code or PFAM ID match a query. In addition, the coordinates and a title or background coloring can be added, and gene-associated quantitative data can be represented by either (bar graph like-) spikes or by a red-to-green gradient. The categorical data have been linked to fixed colors. Map additions are tracked in the 'On display'-panel, in which added rings can also be removed. Users can upload three types of data in the 'Data import'-menu: quantitative data (e.g., expression ratios), the predicted position of sequence elements (e.g., regulatory elements) and/or custom color schemes to designate any genomic region in the genome map. Different genomes can be included in a single circular map, but as synteny is not determined this feature should only be used with very closely related genomes.

Analysis and visualization of sequence elements

An integrated view of genome-wide experimental data and the predicted location of particular regulatory elements can be very allusive in the analysis of transcriptional networks (as illustrated in Fig.1 and supplementary file 1). The position of any particular genomic element with respect to the location and orientation of the surrounding genes can hint at the biological role of that element. CiVi generates plots in the 'Elements and genomic context'-panel for each uploaded set of sequence elements, in which both the distance to the neighboring genes and the orientation with respect to the genes is summarised (Fig 1B). Similarly, the biological role of a particular element may be derived from the functional characteristics of the gene context. The user can download the positions and the annotation data for the gene context for subsequent analysis using the 'Generate context table'-link.

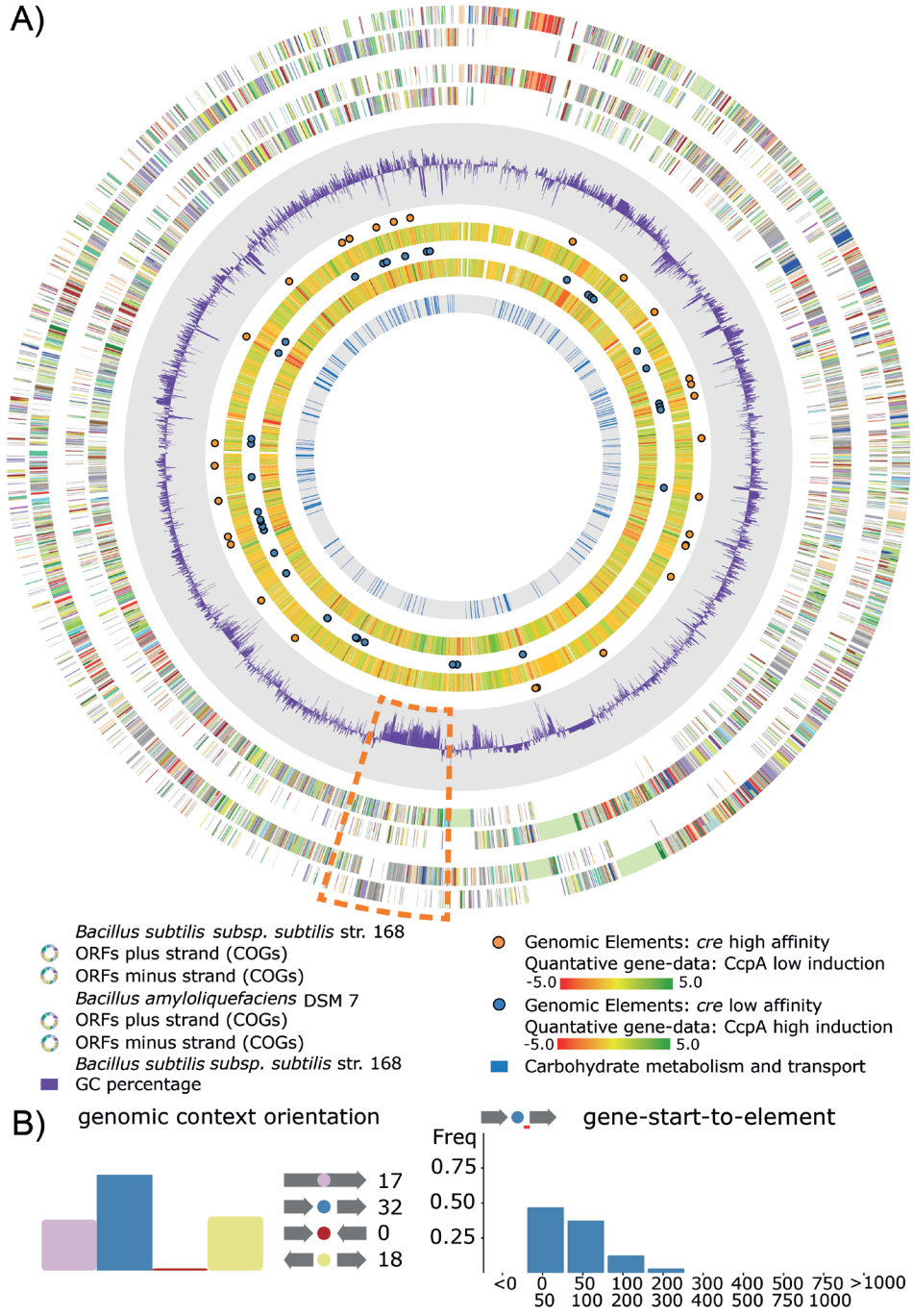

A)
Bacillus subtilis subsp. subtilis str. 168
ORFs plus strand (COGs)
ORFs minus strand (COGs)
Bacillus amyloliquefaciens DSM 7
ORFs plus strand (COGs)
ORFs minus strand (COGs)
Bacillus subtilis subsp. subtilis str. 168
GC percentage
Genomic Elements: cre high affinity
Quantative gene-data: CcpA low induction
-5.0
5.0
Genomic Elements: cre low affinity
Quantative gene-data: CcpA high induction
-5.0
5.0
Carbohydrate metabolism and transport
B)
genomic context orientation
17
32
0
18
gene-start-to-element
Freq
0.75
0.50
0.25
<0
0 50
50 100
100 200
200 300
300 400
400 500
500 750
750 1000
>1000

Figure 1. Application of CiVi in the analysis of genome structure and gene regulation. A) A comparative visualization of the genomes of the closely related *Bacillus* species *B. subtilis* str. 168 (genes and COG annotation on outer two rings) and *B. amyloliquefaciens* DSM 7 (next two rings) shows that the former genome contains a region that seems to have been inserted (indicated by orange box) and which is clearly associated with a deviating GC percentage (5th ring). In *Bacillus* species carbohydrate uptake and metabolism is governed by a phenomenon called 'carbon catabolite repression' which is mediated by CcpA. CcpA binds to a characteristic DNA operator sequence called *cre*, located just upstream of the regulated gene(s). Variability between the operator sequences was assumed to cause variability in the response depending on the level of repression (Francke et al. 2008). The postulated operator dependent variable response was tested experimentally and confirmed (Marciniak et. al. 2012). All of the characteristic properties of CcpA mediated regulation are apparent in an integrative visualization of the transcription factor binding site predictions (6th and 8th ring), gene expression data (7th and 9th ring) and the gene annotation (COG category 'carbohydrate transport and metabolism'; 10th ring). **B)** The provided analysis of the genomic context of the predicted CcpA binding sites confirms the validity of the input motif; as it is mainly found upstream of genes and close to the translation start.

Implementation

CiVi is a web-application developed using a combination of python, javascript, MySQL and SVG. CiVi was implemented as a single page application; the front-end makes server side calls through Jquery and AJAX and receives a response from the server. The maps can be downloaded in SVG, PNG or PDF format, where conversions are done using 'Batik Rasterizer' The maps in SVG-format can be edited in programs such as Adobe Illustrator. CiVi is operable in Firefox, Chrome and Internet Explorer.

Conclusion

CiVi is a versatile and easy-to-use web-application to create custom circular genome maps. It provides a visual integration of publicly available genomic data and additional provided data, the latter including e.g., gene expression data and genomic elements. The functional analysis of latter elements is aided by the characterization of their genomic context, a feature that is unique to CiVi

Supporting information

Supplementary data is freely available online:
https://figshare.com/s/c9b1b9766fa78477c55c

References

Alikhan, N.F., Petty, N. K., Zakour, N.L.B., and Beatson, S.A. (2011). BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC genomics* **12**, 402.

Carver, T., Thomson, N., Bleasby, A., Berriman, M., and Parkhill, J. (2009) DNAPlotter: circular and linear interactive genome visualization. *Bioinformatics* **25**, 119-120.

Conant, G.C. and Wolfe, K.H. (2008). GenomeVx: simple web-based creation of editable circular chromosome maps. *Bioinformatics* **24**, 861-862.

Crabtree, J., Agrawal, S., Mahurkar, A., Myers, G.S., Rasko, D.A., and White (2014) Circleator: flexible circular visualization of genome-associated data with BioPerl and SVG. *Bioinformatics* **30**, 3125-3127.

Francke, C., Kerkhoven, R., Wels, M., and Siezen, R.J. (2008). A generic approach to identify Transcription Factor-specific operator motifs; Inferences for LacI-family mediated regulation in *Lactobacillus plantarum WCFS1*. *BMC Genomics* **9**, 145.

Grant, J.R., Arantes, A.S., and Stothard, P. (2012). Comparing thousands of circular genomes using the CGView Comparison Tool. *BMC genomics* **13**, 202.

Hallin, P.F., Stærfeldt, H.H., Rotenberg, E., Binnewies, T.T., Benham, C.J., and Ussery, D. W (2009). GeneWiz browser: an interactive tool for visualizing sequenced chromosomes. *Standards in genomic sciences* **1.2**, 204.

Kerkhoven, R., Van Enckevort, F.H., Boekhorst, J., Molenaar, D., and Siezen, R.J. (2004). Visualization for genomics: the Microbial Genome Viewer. *Bioinformatics*, **20**, 1812-1814.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Marra, M.A. et al. (2009). Circos: an information aesthetic for comparative genomics. *Gen. res.* **19**(9), 1639-1645.

Marciniak, B.C., Pabijaniak, M., de Jong, A., Dűhring, R., Seidel, G., Hillen, W., and Kuipers, O.P. (2012). High- and low-affinity cre boxes for CcpA binding in *Bacillus subtilis* revealed by genome-wide analysis. *BMC genomics* **13**(1), 401.

Nancy, Y.Y., Laird, M.R., Spencer, C., & Brinkman, F.S. (2010). PSORTdb—an expanded, auto-updated, user-friendly protein subcellular localization database for Bacteria and Archaea. *Nucleic acids res.* **39**, D241-D244.

Chapter 3

Identification and global characterization of repeated sequences in prokaryotic genomes

Lex Overmars, Sacha A.F.T. van Hijum, Roland J. Siezen, Christof Francke

Abstract

Prokaryotic genomes have high protein-coding densities and are relatively compact. Nonetheless, the intergenic regions of these genomes are packed with various elements including sequences that are repeated throughout the chromosome. In eukaryotes, these sequences have been considered to incorporate both structural and functional roles. and likewise in prokaryotes.

In this study we identified abundant repeated sequences with a length of 12 nucleotides (i.e. dodecamers) in the intergenic regions of 1516 prokaryotic chromosomal sequences. A separate search for the identified repeated sequences in the intergenic regions was conducted for all complete genomes. The search resulted in 583 repeated sequences with an occurrence of 40 times or higher in one or more individual genomes. To aid the functional characterization we formulated a strategy based on the distribution profile of the repeated sequences. The profile consists of: i) the taxonomic distribution, ii) the density and genome-wide distribution, and iii) the distribution with respect to the local gene organization.

A group of sequences related to the sequence AAAAATAAAAAA were the most abundant repeated sequences that we identified in this study. Using the distribution profiles we could characterize various repeats, such as the DNA Uptake Sequence (DUS) in *Neisseria* spp, the Repetitive Extragenic Palindromic sequences (REPs) in *Gammaproteobacteria* and the highly repetitive motif (HRM) in *Lactococcus Lactis*. We also identified repeated sequences that can be potentially used to profile different species and describe an example sequence that allows to discriminate between *Lactococcus lactis* strains.

Here we present a set of repeated sequences and their corresponding distribution profiles, which include a summary of their taxonomic distribution, their genomic distribution and their distribution with respect to the local gene organization. Tailored and in-depth analysis is required to link a biological role to an individual repeated sequence. The generated data provide a global but valuable foundation for these analyses.

Introduction

The size of prokaryotic genomes varies greatly. Thus far, the smallest sequenced bacterial genome, the genome of the endosymbiont *Tettigades undata*, is about 135 kb (Van Leuven et al., 2014), whereas the largest, the genome of *Sorangium cellulosum,* is over 14 Mbp (Han et al., 2013). An important factor that explains this range is the large variation in metabolic complexity that prokaryotes exhibit. In addition, prokaryotic genomes markedly vary in 'compactness' (i.e. coding density) (Koonin and Wolf, 2008). The most compactly organized genomes only consist of about 5% (*Thermotoga neapolitana,* genome size 1.88Mb) intergenic sequence regions whereas others consist of up to 50% (*Sodalis glossinidius,* genome size 4.17Mb) intergenic regions (Koressaar and Remm, 2012). The median of the percentage of intergenic sequence regions per genome was estimated to be 12% in the 613 prokaryotic species analyzed by (Koressaar and Remm, 2012).

The intergenic regions contain sequence elements in prokaryotes, like in eukaryotes, and an interesting group of those elements are formed by repeated sequences (Delihas, 2007). The repeated sequences may be found directly following each other (referred to as 'repeats'), but are also found as individual sequences distributed over the genome. Prokaryotic repeats have been divided by (Achaz et al., 2002) into two subclasses: i) 'low complexity repeats' (typically ranging from mononucleotide to pentanucleotide in size); and ii) 'long repeats'. In this simple classification, 'long repeats' include transposable elements (IS), mini-satellites (repeated units in the range 6–10 bp, spanning hundreds of base pairs), tandem repeats and spaced repeats.

Repeated sequences can arise by different processes such as duplication, horizontal gene transfer, transposition, and replicon fusion (Achaz et al., 2002). Repeated sequences are targeted by recombination and thereby increase the rates of rearrangement, amplification and deletion of genomic material (Treangen et al., 2009). Thereby, they increase genome plasticity and are claimed to play an important role in prokaryotic lifestyle (Hernández-Salmerón et al., 2013; Rocha et al., 1999).

Various families of 'long' repeats, such as CRISPRs (Jansen et al., 2002), MITEs, IS elements (Mahillon and Chandler, 1998), ERICs (Hulton et al., 1991) and REPs (Higgins et al., 1982), have been characterized in more detail. CRISPRs (Clustered Regularly Interspaced Short Palindromic Repeats) are a distinct family of repeated sequences found widely distributed in *Bacteria* and *Archaea* (Sorek et al., 2008). They provide acquired immunity against foreign

genetic elements (Makarova et al., 2006). CRISPRs usually consist of highly conserved DNA sequences, typically 21 to 48 bp in size, that are repeated up to 250 times (Grissa et al., 2007). The repeated sequences are interspaced by distinct sequences of similar length (denoted as spacers), usually 20 to 58 bp. These 'spacers' relate to the genome sequence of distinct viruses and provide the host with the appropriate immune response.

Repetitive extragenic palindromic (REP) sequences form another example of relatively well-characterized repeated sequences. They consist of imperfect palindromic DNA sequences of around 35 bp long and are often found in clusters (called bacterial interspersed mosaic elements; BIMEs (Gilson et al., 1991)), predominantly in the genomes of *Gammaproteobacteria*. There have been several studies that linked REPs to key biological roles in the cell since their discovery three decades ago (Higgins et al., 1982). The proposed functions included mRNA stabilization (Newbury et al., 1987), and (specific-) REP sequences were also shown to serve as binding sites for DNA polymerase and IHF (Gilson et al., 1990). Species-specific REPs have been identified in human pathogens such as *Escherichia coli*, *Salmonella enterica*, *Neisseria meningitidis*, *Mycobacterium tuberculosis*, *Rickettsia conorii* and *Pseudomonas aeruginosa*, the plant pathogen *Agrobacterium tumefaciens* and the soil bacteria *Deinococcus radiodurans*, *Pseudomonas putida* and *Sinorhizobium meliloti* (Tobes and Ramos, 2005). REPs were found almost exclusively in the intergenic region between co-oriented and convergent gene-pairs (Tobes and Ramos, 2005).

Another well-characterized repeated sequence is the Enterobacterial Repetitive Intergenic Consensus (ERIC) sequence. ERICs resemble REPs as they are also located in non-coding regions of the chromosome and they include a conserved inverted repeat thereby encoding a potential stem-loop structure (Hulton et al., 1991).

Transposable elements (TEs) are an important source of repeated sequences in bacterial and archaeal genomes (Touchon and Rocha, 2007). The most abundant TEs among Bacteria and Archaea are Insertion Sequences or IS elements. The length of IS elements typically ranges from 0.7 to 3.5 kbp. They contain a transposase gene and are flanked by imperfect terminal repeat sequences (Mahillon and Chandler, 1998). IS elements can occupy a significant fraction of a prokaryotic genome, going up to 40% in the genome of the bacterium *Orientia tsutsugamushi* (Cho et al., 2007). Yet, their occurrence varies greatly between different species and even between closely related genomes (Filée et al., 2007). Non-autonomous miniature inverted-repeat transposable elements (MITEs) are a special type of transposable elements

that were first discovered in plants (Feschotte and Mouchès, 2000). MITEs are highly abundant in many eukaryotic genomes. However, they are also present in bacterial genomes, where they are primarily found in intergenic regions, but also present intragenically (Ogata et al., 2000). MITEs share characteristics with insertion sequences (ISs) as they contain terminal inverted repeats (TIRs), and are flanked by target site duplications, which in turn consist of direct repeats (DRs). However, in contrast to ISs, MITEs do not contain a transposase gene (Delihas, 2007).

Repeated sequences have been used as targets for DNA profiling of prokaryotes. For instance, REP-PCR is a type of PCR that targets the repeated sequences in prokaryotic genomes using specific primers complementary to REP and/or ERIC sequences. REP-PCR was shown to be a reproducible fingerprinting technique, applicable to a wide range of prokaryotic phyla (Gevers et al., 2001; Rademaker et al., 2004; Versalovic et al., 1991).

We decided to analyze the presence of repeated sequences in the bacterial and archaeal genomes available through NCBI to evaluate the possibility of using other marker sequences for profiling, and moreover, to uncover potential new sequences of interest. We focused on more abundant repeated sequences; i.e. sequences that occurred at least 40 times in one of the selected genome sequences. We also defined a genomic location based strategy to help characterize the biological role/molecular function of highly abundant repeated sequences. We identified all repeated sequences of 12 nucleotides (dodecamers) (the lowest number needed to reduce the random occurrence of given sequence in any bacterial genome to <1) in the 1516 individual bacterial and archaeal genomes that we acquired from the NCBI RefSeq database and found 583 characteristic 12-nucleotide sequences. Subsequently, the occurrence of these sequences was determined for each genome.

A combinatorial analysis of the distribution over the taxonomy, the distribution over the chromosome and the distribution with respect to local gene organization appeared an effective strategy to create a distribution profile for every repeated sequence. We were thus able to characterize various exemplary repeated sequences, such as the Repetitive Extragenic Palindromic sequences (REP) in *Gammaproteobacteria* (Higgins et al., 1982), the Highly repetitive motif (HRM) in *Lactococcus lactis* (Mrázek et al., 2002) and the DNA uptake sequence (DUS) within the genus Neisseria (Duffin and Seifert, 2010). Moreover, we identified many others that potentially represent different functions.

Material and methods

Genome sequences and annotation

The genome sequences and annotation of publicly available microbial genomes were obtained from the FTP server of NCBI RefSeq (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/; Pruitt et al., 2012). The data was accessed on April 23 2012 and all completed genomes were included. The dataset generated in this way consisted of 1516 chromosomal sequences (bacterial as well as archaeal; no plasmids). The taxonomic annotation: super kingdom, phylum class, order, family, genus and species, was collected for each genome sequence using the NCBI taxonomy database (ftp://ftp.ncbi.nlm.nih.gov/pub/taxonomy/).

De novo identification of overrepresented repeats

The annotated intergenic regions were extracted from each genome sequence and grouped based on local gene organization: i) co-oriented- [→..→], ii) divergent- [←..→] and iii) convergent- [→..←] intergenic region. Each of these groups were then subjected to a separate *de novo* motif discovery search using MEME 4.8.1 (Bailey et al., 2006) to identify fully conserved sequences with a length of 12 bp. The following search parameters were used: the number of motifs was set to 5 per region and the site selection was set to any number of repetitions. A fixed motif sequence length of 12 bp was chosen to ensure a high recovery (small length) and at the same time to decrease the probability of random occurrences of the identified sequences in any bacterial genome to <1 (random occurrence = one 12bp sequence per 16 Mbp; the average genome length of the sequenced bacterial genomes was around 3.7 Mbp (Lagesen et al., 2010)).

Analysis of the identified repeats

To include only highly abundant repeated sequences, motifs with at least 40 exact matches (i.e. completely identical) in a single genome were included for further analysis. The similar motif search (SMS) procedure (Francke et al., 2011) was used to search the available bacterial and archaeal genome sequences for the occurrence of each of the identified repeated sequences. Each hit was categorized based on the local orientation of the adjacent genes: i) intragenic [-..→], or ii) co-oriented- [→..→], iii) divergent- [←..→] and iv) convergent- [→..←] intergenic. The results were summarized in a matrix in which the absolute number of hits per repeat per genome is shown (Table S1; https://figshare.com/s/f707d70e804d07179e99). For each of the repeated sequences an average distribution of location with respect to the neighboring

genes was calculated. For the calculation only genomes with >10 copies were included. The repeated sequences were clustered on the basis of the average distribution (method: Ward's method; distance measure: Euclidean).
The GC%-dependent expected number of copies per 1 million basepairs was calculated as follows: *((0.5*GC)^GC_count)*((0.5*(1-GC))^AT_count)*1000000*, where GC was the GC percentage of the genome, GC_count the number G or C bases in the repeated sequence and AT_count the number of A or T bases in the repeated sequence.

Visualization of the genome-wide distribution of the identified repeats

The genome-wide distribution of any particular conserved repeated sequence was visualized using CiVI (this thesis Chapter 2b; (Overmars et al., 2015a)). CiVI was also used to visualize the distribution of the repeated sequences with respect to the local orientation of the genes. Depending on the orientation of the neighboring genes, i.e. the type of genomic region the elements are located in, the following distances were analyzed: i) intragenic hits: gene-start-to-12 bp repeated sequence and gene-stop-to-12 bp repeated sequence; ii) co-oriented hits: gene-start-to-12 bp repeated sequence and gene-stop-to-12 bp repeated sequence; iii) divergent hits: gene-start-to-12 bp repeated sequence; and iv) convergent hits: gene-stop-to-12 bp repeated sequence.

Results and Discussion

Analysis of overrepresented repeats and their taxonomic distribution

Overrepresented fully conserved sequences of length 12 bp were identified in the intergenic regions of the 1516 complete prokaryotic genome sequences in the NCBI database with MEME (Bailey et al., 2006) (see methods). Of the 22755 recovered repeated sequences, 583 occurred at least 40 times within one of the genome sequences in which they were identified. The 583 abundant 12 bp repeated sequences were subsequently searched in all selected genomes using a Similar Motif Search (see methods) to determine their taxonomic distribution (Table S1; https://figshare.com/s/f707d70e804d07179e99).

The genome-wide searches resulted in 1165268 hits in total within the 1516 genomes (Fig. 1 and Table S1). The most abundant sequence (i.e. most occurrences within the 1516 chromosomes) appeared the A-rich sequence AAAAATAAAAAA (1st most abundant; see Table 1 and Table S1). We found a total 37574 occurrences of this repeat in 1516 genomes and identified 801 genomes carrying more than 10 copies. The occurrence of the adenine-rich repeats AAAAAATAAAAA (2nd most abundant) and AAAATAAAAAAA (3rd

most abundant) was mostly similar, in line with the single nucleotide shift in sequence. The taxonomic distribution and occurrence of the repeated sequences AAAATAATAAAA (4th most abundant) and TAAATTTAAAAA was also strongly correlated to that of AAAAATAAAAAA (Fig. 1; indicated in red). Moreover, a similar taxonomic distribution was found for the abundant repeat TTTTCTTTTTTT (6th most abundant). The fifth and seventh most abundant repeated sequence were CG-rich. The repeated sequence CGCGCGCCGCGC was found 18694 times and CGCCTGCGCGGC was found 13495 times. Interestingly, the taxonomic distribution of these CG-rich repeats seemed inversely related to the distribution of the A-rich repeats.

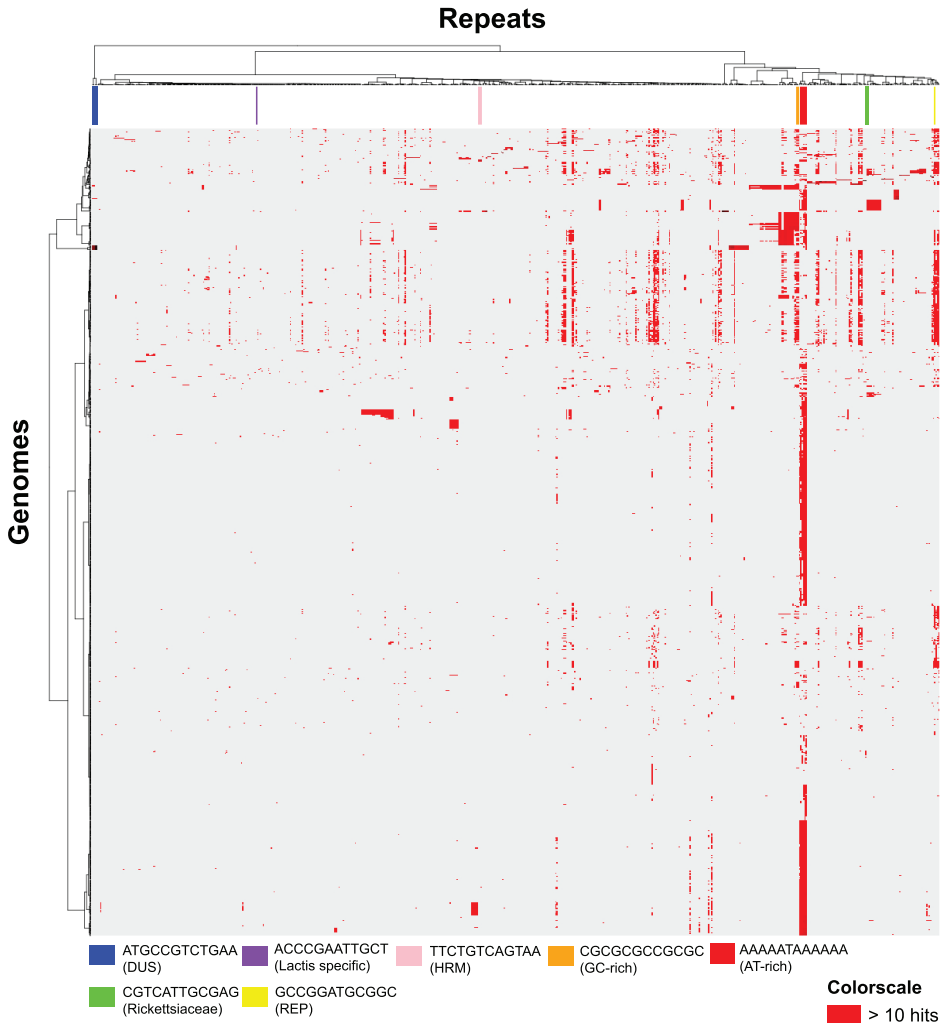


Figure 1. Heatmap of the occurrence of the 626 identified repeated sequences. Red: occurrence of > 10 within an individual chromosome. Hierarchical clustering (method: Ward's method; distance measure: Euclidean) is done both on the x-axis (repeats) and the y-axis (genomes).

We found that the taxonomic distribution of most of the identified repeated sequences was more restricted (Table 1 and Table S1). Some, such as the A-rich repeated sequences, appeared characteristic for a larger group of species, whereas others were highly species- or even strain-specific. The phylum of *Proteobacteria* included relatively the most genomes carrying repeated sequences. Many repeated sequences were also identified in the bacterial phyla *Actinobacteria*, *Bacteroidetes*, *Cyanobacteria*, *Firmicutes* and *Tenericutes*, and the archaeal phylum *Euryarchaeota* (Fig. 2 and Table S1).

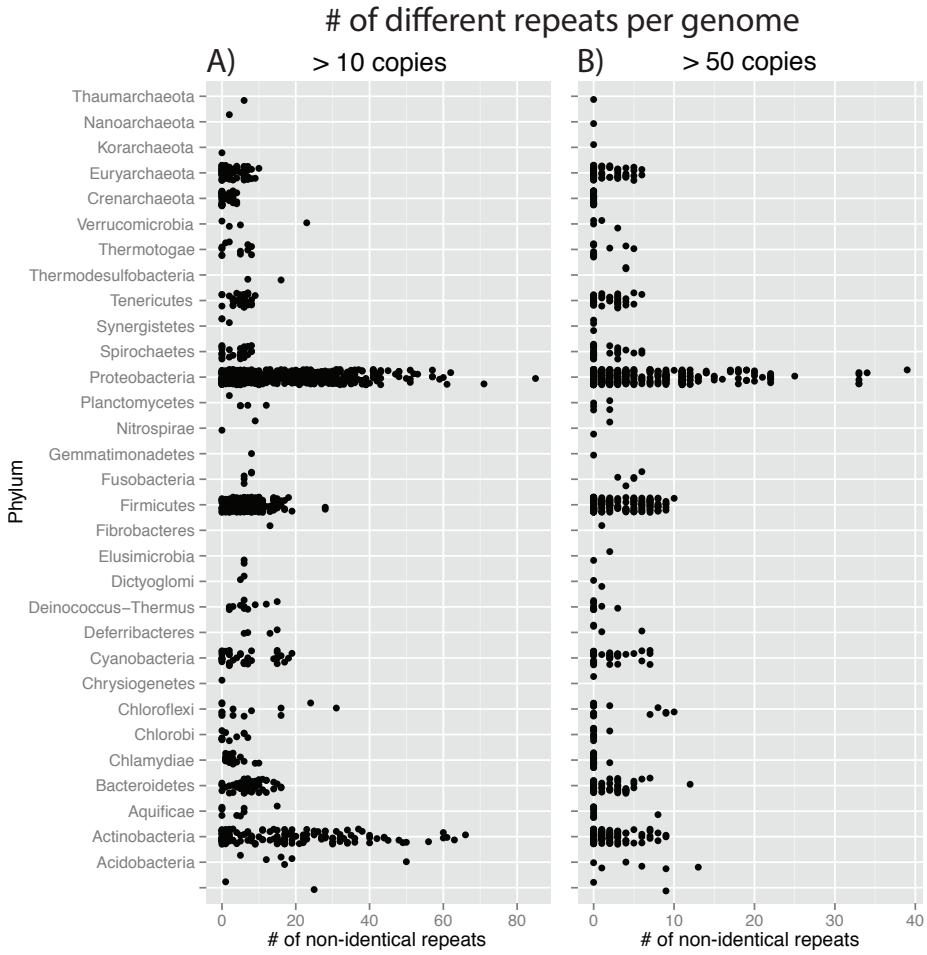


Figure 2. The abundance of repeated sequences per phylum. A) Each dot represents an individual genome (grouped by phylum), where the x-axis indicates the number of (non-identical) repeats with at least 10 copies. **B)** Each dot represents an individual genome (grouped by phylum), where the x-axis indicates the number of (non-identical) repeats with at least 50 copies.

We identified a remarkably high number of species/strain-specific repeated sequences (Table 2 and Table S1). The highest abundance of a species/strain-specific repeated sequence was found in the genome sequence of *Neisseria lactamica* 020-06, in which the sequence ATGCCGTCTGAA was present in 1723 copies**.** Other species with remarkably high 'copy numbers' included *Krokinobacter* sp. 4H-3-7-5, *Microcystis aeruginosa* NIES-843 and species of the family of *Pseudomonadaceae* (Table 2).

Table 1. Top 10 of the most abundant repeated sequences. Abundance is expressed as the total number of copies in the 1516 analyzed genomes. Genomes>10 represents the number of genomes in which > 10 copies of the repeated sequence were present, Phyla >10 represents the number of phyla in which at least one individual genome contained the repeated sequence in >10 copies. Max indicates the maximum number of copies within an individual genome.

Repeat	Total	Genomes>10	Phyla>10	Max
AAAAATAAAAAA*	37574	801	24	305
CGCGCGCCGCGC	18694	326	11	841
TTTTCTTTTTTT	16625	484	22	183
CGCCTGCGCGGC	13495	328	12	227
TTCAGACGGCAT	12443	10	1	1723
TAAATTTAAAAA	12262	349	17	406
CGACGACGCCGC	11615	220	11	471
CAGCGCCGCGCG	11393	293	11	185
CGCTGGCCGGCA	10142	250	9	596
GCCGCATCCGGC	9551	218	7	181

* Only the most abundant sequence was included in case sequences clearly correspond in terms of sequence similarity and occurrence (e.g. AAAAAATAAAAA (33933), AAAATAAAAAAA (28703), and AAAATAATAAAA (19335))

Table 2. Top 10 of the most abundant repeated sequences within a single genome. Max indicates the maximum number of copies within an individual genome. The corresponding species is indicated (Max Abundance Species). Total represents the total number of copies within the 1516 analyzed genomes, whereas Genomes > 10 represents the number of genomes with > 10 copies of the repeated sequence.

Repeat	Max	Total	Genomes>10	Max. Abundance Species
ATGCCGTCTGAA	1723	12443	10	*Neisseria lactamica* 020-06
CGCTTTCGCGAA	1452	2811	5	*Krokinobacter* sp. 4H-3-7-5
CACCCCACACCC	1386	2269	12	*Microcystis aeruginosa* NIES-843
GCTTGCTCGCGA	1296	3574	11	*Pseudomona brassicacearum* NFM421
CCGCTCCCACAG	1262	3377	16	*Pseudomonas putida* S16
ACTGATAACTGA	1050	1887	8	*Microcystis aeruginosa* NIES-843
GGCGATCGCCAA	944	6678	136	*Synechococcus* sp. PCC 7002
AATTCTTAATTC	944	1924	4	*Chitinophaga pinensis* DSM 2588
GCTCGGCCCTGC	908	2473	23	*Saprospira grandis* str. Lewin
GCTCGGAATGAC	891	1482	4	*Roseiflexus castenholzii* DSM 13941

* Only the most abundant sequence was included in case sequences clearly corresponded in terms of sequence similarity and occurrence (e.g. CACCCCACACCC was selected over CCCCACACCCCA; GCTTGCTCGCGA was selected over AGCTTGCTCGCG and ATCGCGAGCAAG; CCGCTCCCACAG was selected over CCCGCTCCCACA)

To characterize the repeated sequences further we determined the organization of the repeated sequences with respect to the orientation of the neighboring genes (Table S1). The 'seed search' for repeated sequences was performed on non-coding sequences to prevent spurious hits from particular protein sequence associated patterns. Nevertheless, some of the most-abundant repeated sequences that we identified, such as the A-rich sequences, were found mainly in coding regions (Table S1; https://figshare.com/s/f707d70e804d07179e99). Actually, only few repeated sequences were found to be specific for the intergenic region. We averaged the relative occurrence in the intragenic and co-oriented-/ divergent-/ convergent –intergenic regions (for each orientation using the fraction of the total per genome) for each repeated sequence using the genomes in which they occurred > 10 times. We performed a hierarchical clustering of the repeated sequences based on the average occurrences (Fig. 3). A distinct cluster (indicated in green in Fig. 3) consisted of the repeated sequences that were primarily found in the intragenic regions (average fraction >0.6). The other clusters included the repeats that were found within co-oriented intergenic regions and could be divided into i) mainly co-oriented, ii) co-oriented + intragenic, iii) co-oriented + convergent and iv) co-oriented + convergent + intragenic (Fig. 3).

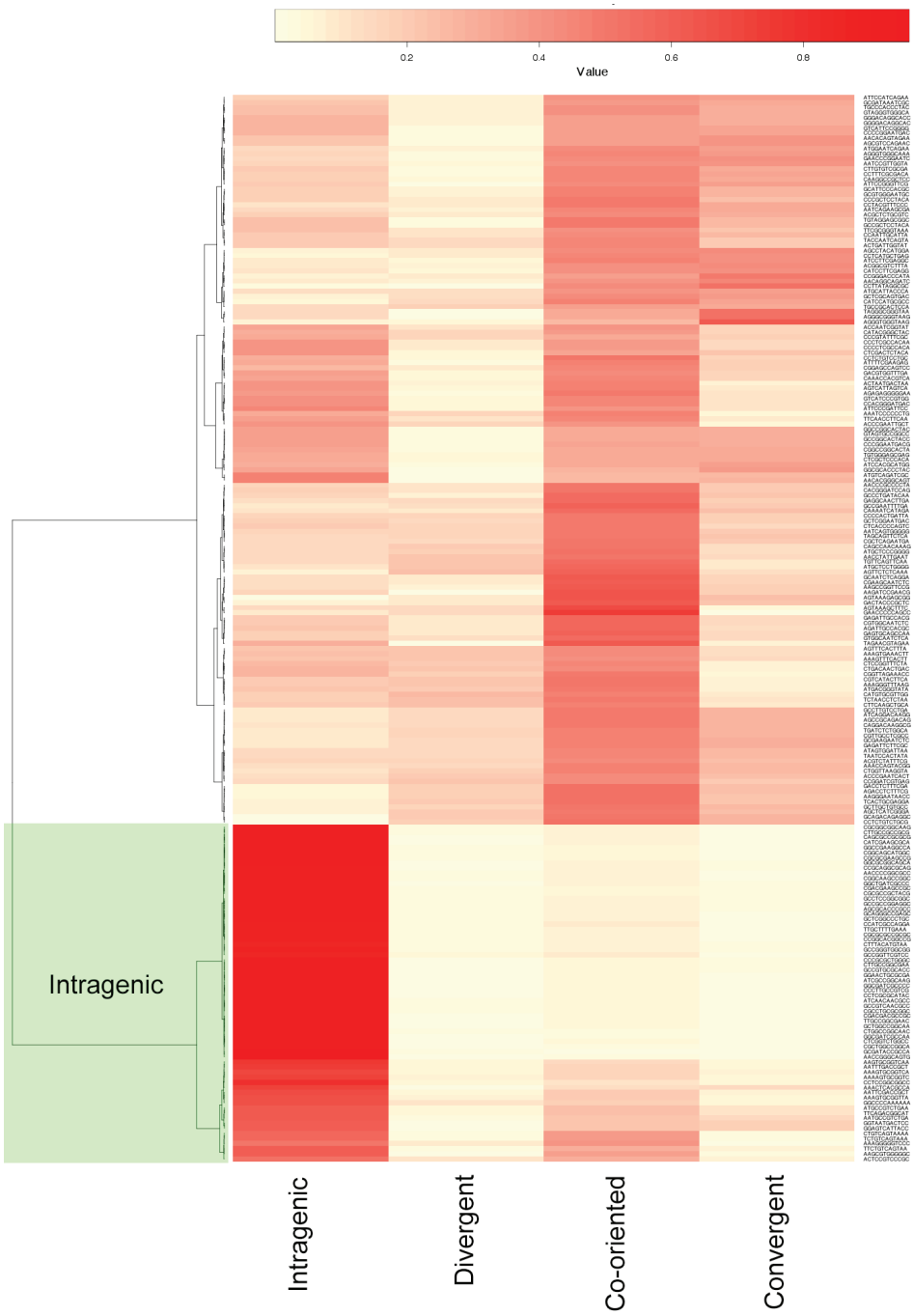


Figure 3. Heatmap of the average distribution of the orientation of the neighboring genes per repeated sequence. A repeat was included when i) the repeated sequence occurred in 10 or more copies in 2 or more genomes and ii) the square root of the average of the variances of the gene organization ratios was < 0.01. The latter was done to select for repeats with a consistent profile.

Below we discuss the taxonomic distribution, genomic distribution and distribution with respect to local gene organization for various repeated sequences. The selected sequences are used to illustrate the application of the formulated strategy to profile newly identified sequence elements in terms of location. Assigning a detailed function to individual sequences was beyond the scope of the presented work because that would require an analysis tailored specifically for each repeated sequence. It would require additional analysis of the annotated function of the gene context., and in many cases even an annotation of the gene context. The selected examples include some of the most abundant repeated sequences, such as the A-rich repeated sequence and the GC-rich repeated sequence (Table 1), but also the most abundant repeated sequence within a single genome, ATGCCGTCTGAA. In addition, the sequence GCCGCATCCGGC is discussed, a repeat within the top 10 abundant repeated sequences with a distinct distribution with respect to the local gene organization. Finally, we highlight various repeats that were found to be species- or strain- specific.

AAAAATAAAAAA, TTTTCTTTTTTT and GC-rich repeats

AAAAATAAAAAA was the overall most abundant repeated sequence that we identified (Table 1). In fact, its sequence is very similar to that of the three other highly abundant repeats, AAAAAATAAAAA (33853), AAAATAAAAAAA (28644) and AAAATAATAAAA (19274) (Fig. 1; indicated in red). The AAAAATAAAAAA repeat was found in most phyla (Fig. 4A). It occurs in large numbers (>100) in one or more genomes within the phyla *Bacteroidetes*, *Cyanobacteria*, *Euryarchaeota*, *Firmicutes*, *Fusobacteria*, *Proteobacteria*, *Spirochaetes*, *Tenericutes* and *Thermotogae*. We observed that the occurrence of AAAAATAAAAAA was negatively correlated to GC% (Fig. 4B; Spearman's rho -0.88). However, the occurrence strongly deviates from random independent of the GC%, indicating that the sequence will probably serve a similar purpose for all genomes.

The AAAAATAAAAAA sequence was found in all four possible gene organization contexts (i.e. intragenic, co-oriented-, divergent- and convergent intergenic region), although there was a clear overrepresentation of copies in the coding sequence (i.e. intragenic) and in co-oriented intergenic regions (Table S1). Within the genomes that contain at least 30 copies (375 genomes; 27827 copies), 60.5% of the copies were located in the coding sequence, whereas 25.4% were located between co-oriented genes. Respectively, 8.7% and 5.4% were found within divergent- and convergent intergenic regions. The sequence was found distributed throughout the circular chromosomes. For example, within the genome of *Sebaldella termitidis* ATCC 33386

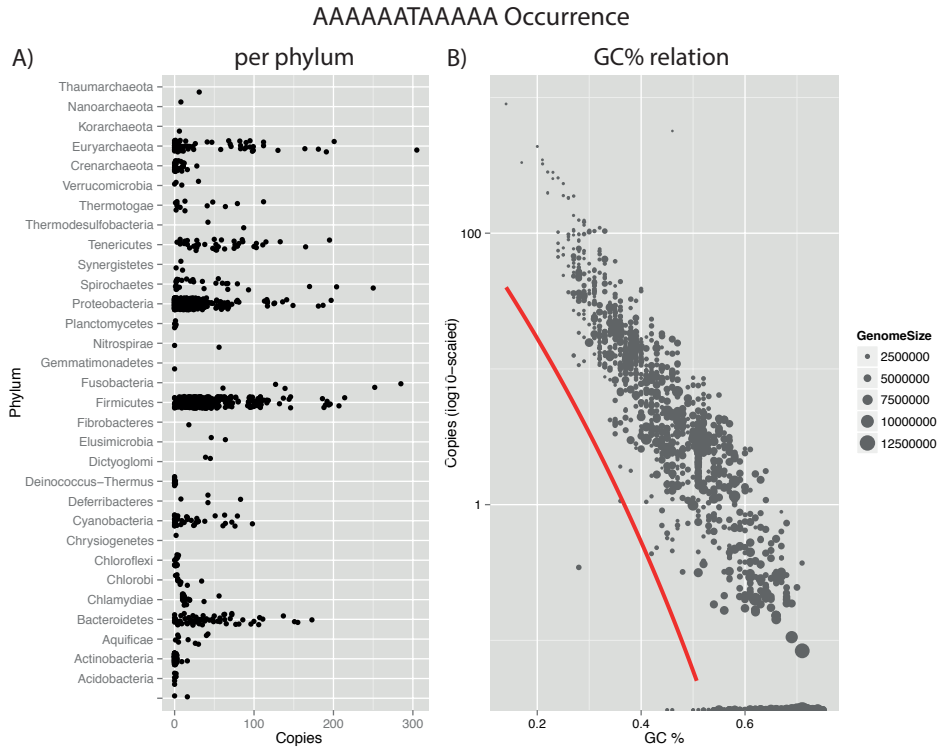


Figure 4. Taxonomic distribution of the repeated sequence AAAAAATAAAAA. A) The number of copies of AAAAAATAAAAA per genome, grouped by phylum. **B)** Scatterplot of normalized occurrences (number of occurrences per 1Mb) versus GC content. The red line indicates the GC dependent expected number of occurrences per 1Mb (see materials and methods); normalized occurrences below 0.05 were not included. Although clearly correlated to AT-content, the observed number of copies of AAAAAATAAAAA was higher than the random expected value in all genomes with at least 1 copy per 1Mb.

(*Fusobacteria*) in which 284 copies were identified uniformly distributed throughout the genome (Fig. 5A). The majority of copies were located in intragenic regions (Fig. 5B). Also no bias was observed in terms of the location with respect to translation start and stop of the neighboring genes (Fig. 5C-G). Similar distributions were observed for most of the other genomes that we inspected. An additional example can be found in Fig. S1, in which the AAAAATAAAAAA distribution within the of genome *Clostridium botulinum* A2 str. Kyoto (*Firmicutes*) is shown. Interestingly, in *Methanobrevibacter ruminantium* M1 (*Euryarchaeota*), the genome in which most copies of the AAAAATAAAAAA sequence were identified, a different positional bias was observed (Fig. S2). Although the 305 copies were distributed across the genome (Fig. S2A), the vast majority of the copies were located within the annotated intergenic regions. These intergenic copies were mostly located closely to the upstream coding sequence. A possible explanation for this positional deviation (i.e. intergenic presence versus intragenic presence)

could be an incorrect annotation of coding sequence, with a bias towards ORFs that are too short. However, the determined TIS annotation score ((Overmars et al., 2015b); this thesis chapter 3) of *Methanobrevibacter ruminantium* M1 was high (0.92), indicating that the genome-wide TIS annotation is probably correct and the sequence AAAAATAAAAAA might have a species specific function in *Methanobrevibacter ruminantium* M1.

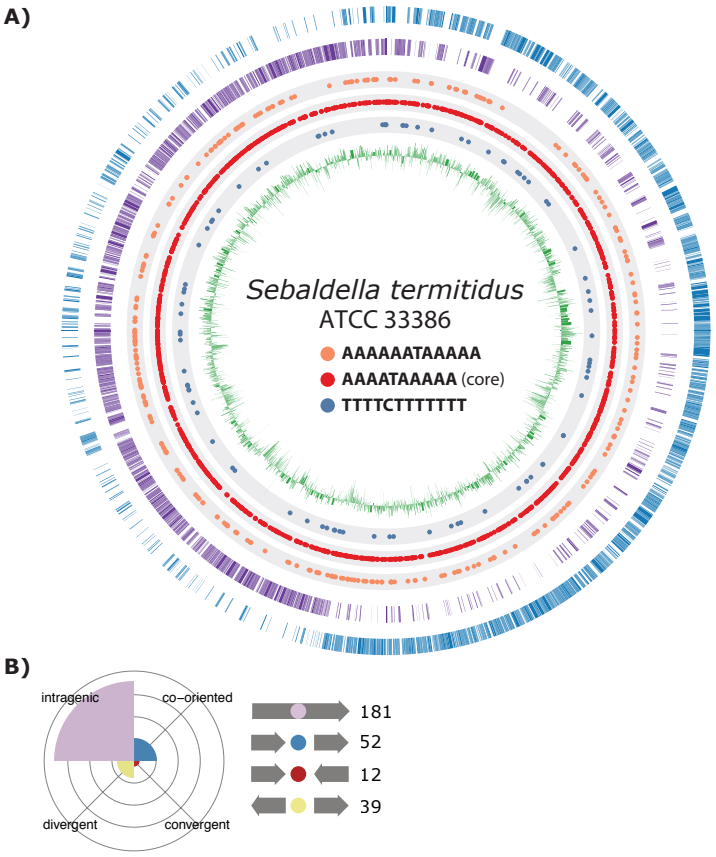


Figure 5 (A-B). The distribution profile of AAAAAATAAAAA in the genome of *Sebaldella termitidis* ATCC 33386. **A)** Circular map of *Sebaldella termitidis* ATCC 33386 with the identified locations of the AAAAAATAAAAA repeat. Blue ring: ORFs on the plus strand, purple ring: ORFs on the minus strand, orange circles: position of the AAAAAATAAAAA sequences, red circles: position of the AAAATAAAAA (core-) sequences, blue circles: position of the TTTTCTTTTTTT sequences and green ring: GC-percentage. **B)** Local gene organization of genes adjacent to the AAAAAATAAAAA sequence.

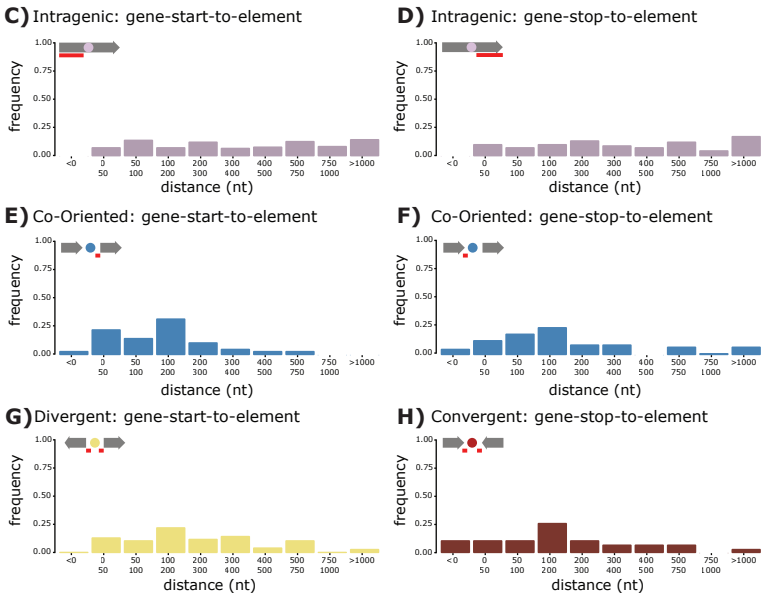


Figure 5 (C-H). The distribution profile of AAAAAATAAAAA in the genome of *Sebaldella termitidis* ATCC 33386. C) and **D)** Distribution of the distance of the intragenic copies to the start and stop codon of the gene in which they are located. **E)** and **F)** Distribution of the distance of copies located in co-oriented intergenic regions to the downstream start codon and the upstream stop codon. **G)** Distribution of the distance of copies located in divergent intergenic regions to the nearest start codon. **H)** Distribution of the distance of copies located in convergent intergenic regions to the nearest stop codon.

The overrepresentation of A-rich repeats was reported before (Davenport and Tümmler, 2010). Davenport and Tümmler suggested that these A-rich repeats could have various functions, which might explain why these sequences are so commonly over-represented. 'A-tracts' were posed to lead to intrinsic DNA-bending (Haran and Mohanty, 2009), but were also reported to be associated to DNA-binding proteins that relate to the packaging of the nucleoid, such as histone like protein HU and nucleoid structuring protein H-NS (Swinger and Rice, 2004). In addition, the A-rich repeats have been related to a localization in the terminal loop of superhelical domains (Haran and Mohanty, 2009). They were also presumed to be associated to transcriptional regulation, which is unlikely given the abundant intragenic presence and dispersed presence in the intergenic regions (binding sites are often concentrated near the translation start). A-tracts could accumulate macroscopic curvature when repeated in tandem with the helical repeat (Tolstorukov et al., 2005).

The taxonomic distribution of the AAAAATAAAAAA sequences correlated with the presence of the TTTTCTTTTT sequences. The overall distribution of TTTTCTTTTT had similarities to those of the A-rich sequences: they were both relatively uniformly distributed throughout the genomes and

their distribution with respect to local gene-organization was highly similar. Within the genomes that contain at least 30 copies (174 genomes; 8015 total copies), a majority (72.2%) of the copies were located in the coding sequence, whereas the intergenic copies were distributed as follows: 17.2% co-oriented, 6.1% divergent and 4.4% convergent (Table S1; https://figshare.com/s/f707d70e804d07179e99). We identified 227 copies of TTTTCTTTTT in the genome of *Sebaldella termitidis* ATCC 33386 (Fig. 5). We did not observe any consistent co-occurrence between TTTTCTTTTT and AAAAATAAAAAA within the genome; neither any overrepresented distance separating TTTTCTTTTT copies was observed (i.e. no helicity was observed).

In many genomes in which the AAAAATAAAAAA-like repeated sequences were absent, high GC-containing repeats such as CGCGCGCCGCGC (18694 copies), CGCCTGCGCGGC (13495 copies), CGACGACGCCGC (11615 copies), CAGCGCCGCGCG (11393 copies), CGCTGGCCGGCA (10142), and some more were found. In fact, the occurrence of the A-rich and T-rich repeats on the one hand, and the GC-rich repeats on the other, seem anti-correlated. For example, the correlation between the occurrence of AAAAATAAAAAA and CGCGCGCCGCGC was -0.66 (Spearman's rho). The CGCGCGCCGCGC repeated sequence was abundant in *Proteobacteria*, *Actinobacteria*, *Deinococcus* and *Chloroflexi* (Fig. 6A). The bias towards an intragenic location of CGCGCGCCGCGC is greater than that of the A-rich and T-rich sequences (Table S1). Within genomes containing at least 30 copies (178 genomes; 14868 copies), 87.9% of the copies were located in the coding sequence, whereas the intergenic copies were distributed as follows: 7.1% co-oriented, 3.2% divergent and 1.9% convergent. The occurrence of GC-rich repeated sequences was clearly correlated with GC% (Fig. 6B; the correlation between CGCCTGCGCGGC and GC% was 0.86). However at the same time, the occurrence strongly deviates from random, independent of the GC%, indicating that the sequence might have a similar functional role in the genomes.

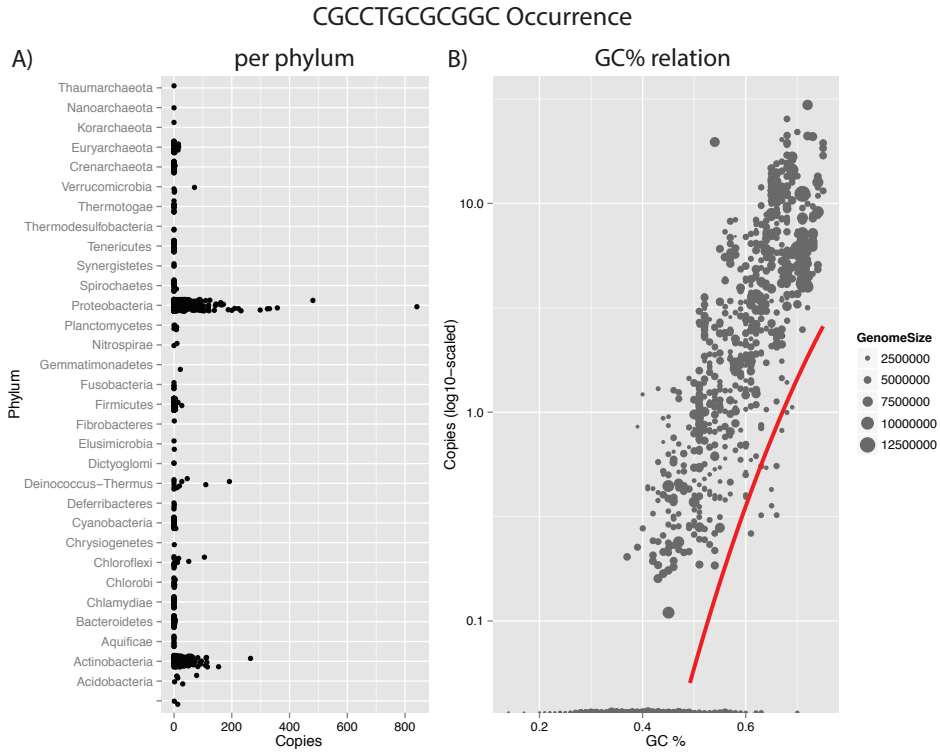


Figure 6. Taxonomic distribution of the repeated sequence CGCGCGCCGCGC. A) The number of copies of CGCGCGCCGCGC per genome, grouped by phylum. **B)** Scatterplot of normalized occurrences (number of occurrences per 1Mb) versus GC content. The red line indicates the GC dependent expected number of occurrences per 1Mb (see materials and methods); values below 0.05 were not included. Although clearly correlated to GC-content, the observed number of copies of CGCGCGCCGCGC was higher than the random expected value in almost all genomes with at least 1 copy per 1Mb.

Repetitive Extragenic Palindromic sequences (REP) in Gammaproteobacteria

The sequence GCCGGATGCGGC and its reverse complement counterpart (GCCGCATCCGGC) were identified 9551 times within the 1516 genomes. The sequences were found mainly within *Proteobacteria* (Fig. 7), but were also found in *Actinobacteria* and to a lesser extent in some *Firmicutes*. In literature, this sequence is described as the Y-type of a palindromic unit denoted as Repetitive Extragenic Palindromic sequence (Gilson et al., 1991). The guanine in the fourth position is a critical base in distinguishing the Y-type from the Z1- and Z2- type (in which the guanine is replaced by a thymine) (Bachellier et al., 1994). The Y-type repeat was most abundant in the genome of *Escherichia coli str.* K-12 substr. MG1655, in which it occurred 181 times (Fig. 8A). The Y-type fragment repeats were located in the intergenic region and were mostly found between convergent- and co-oriented gene-pairs (Fig. 8B). Some copies were positioned intragenically, but all of these copies were

located within 50 nucleotides of the stop codon of the gene in which they were found (Fig. 8C). Other *E. coli* REP related repeats that we have identified included (A)ACGCCTTATCC(G) (AACGCCTTATCC and ACGCCTTATCCG), (CG)CCTTATCCGG(CC), CCGCATCCGGCA and AGGCC(G/T)GATAAG, which are corresponding to the second part of the Y-type REP sequence. The distinct gene-organization profile was described before for REP sequences (Tobes and Ramos, 2005). We found many repeated sequences in the *Gammaproteobacteria* and in other phyla that showed a similar distribution with respect to the neighboring genes (see table S1). The role of these repeats might be comparable to the REP sequences.

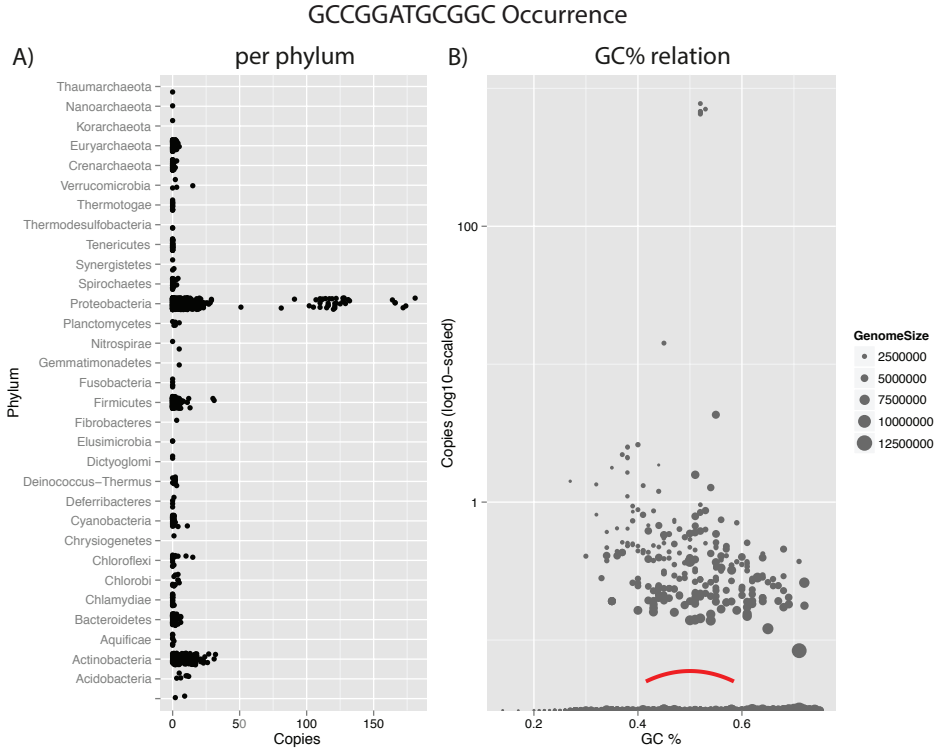


Figure 7. Taxonomic distribution of the repeated sequence GCCGGATGCGGC. A) The number of copies of GCCGGATGCGGC per genome, grouped by phylum. **B)** Scatterplot of normalized occurrences (number of occurrences per 1Mb) versus GC content. The red line indicates the GC dependent expected number of occurrences per 1Mb (see materials and methods); values below 0.05 were not included.

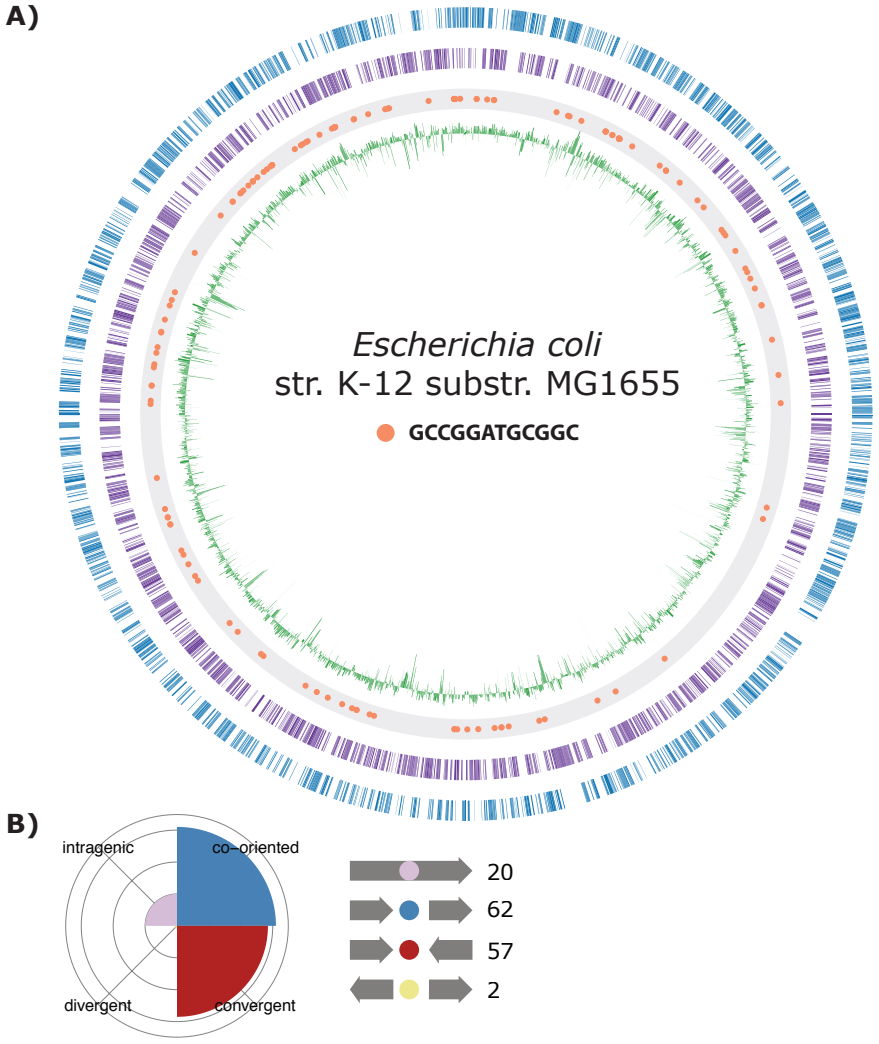


Figure 8 (A-B). The distribution profile of GCCGGATGCGGC in the genome of *Escherichia coli* str. K-12 substr. MG1655. **A)** Circular map of *E. coli* str. K-12 MG1655 with the identified locations of the GCCGGATGCGGC repeat (orange). In addition, the following rings were included: i) ORFs on the plus strand (blue), ii) ORFs on the minus strand (purple) and iii) GC-percentage (green). **B)** Local gene organization of genes adjacent to the GCCGGATGCGGC sequence.

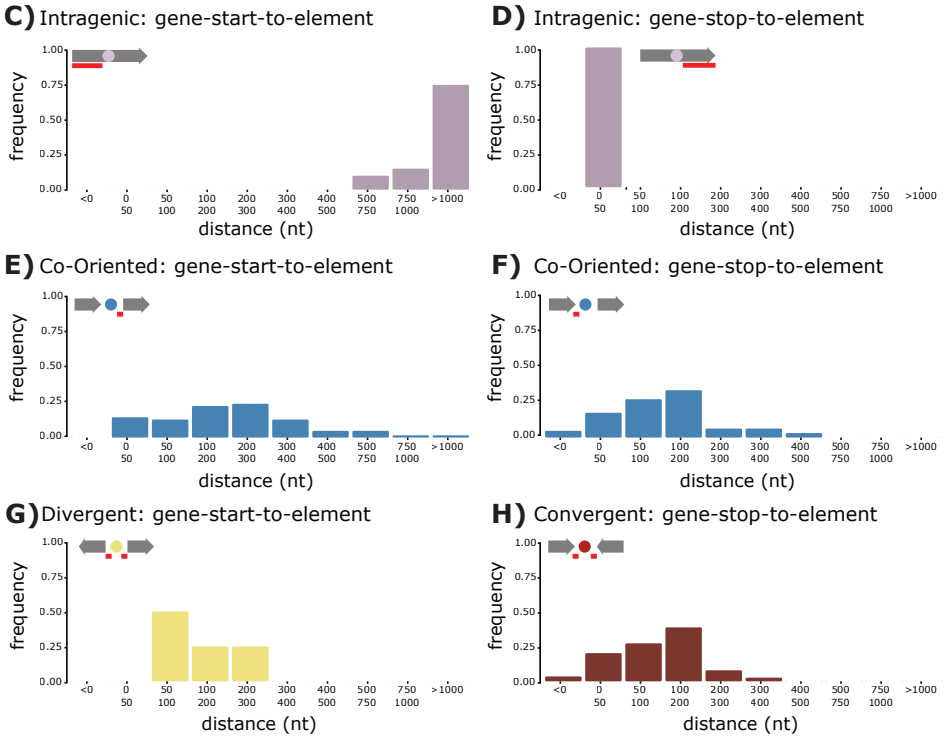


Figure 8 (C-H). The distribution profile of GCCGGATGCGGC in the genome of *Escherichia coli* str. K-12 substr. MG1655. C) and **D)** Distribution of the distance of intragenic copies to the start and stop codon of the gene in which they are located. **E)** and **F)** Distribution of the distance of copies located in co-oriented intergenic regions to the downstream start codon and the upstream stop codon. **G)** Distribution of the distance of copies located in divergent intergenic regions to the nearest start codon of a gene. **H)** Distribution of the distance of copies located in convergent intergenic regions to the nearest stop codon of a gene.

The DNA uptake sequence (DUS) of Neisseria spp.

The sequence ATGCCGTCTGAA was highly abundant within the 8 genomes of the genus *Neisseria*. We identified 1723 hits in the genome of *Neisseria lactamica* 020-06. Also for the other *Neisseria* genomes an occurrence of >1400 was found. Besides its occurrence in the *Neisseria* genus, the repeated sequence was only found in significant numbers in the genome of *Actinobacillus succinogenes* 130Z (33 hits) and the genome of *Citrobacter rodentium* ICC168 (23 hits). The sequence was found evenly distributed over the complete genome of *Neisseria lactamica* 020-06, except for some low-GC regions in which it was absent (Fig. 9A). The sequences were present within the intragenic regions (651 hits), but also in the intergenic regions (772 hits) (Fig. 9B). The ATGCCGTCTGAA sequences positioned between co-oriented genes were almost all located within a range of 50 to 100 nucleotides of the

3' end of the upstream gene (Fig. 9F), whereas the vast majority of intragenic sequences were located within 50 nucleotides of the 5' end of the gene (Fig. 9D). Consistent with this preference, the sequences located within convergent intergenic regions were mostly found in a range of 50 to 100 nucleotides to one of the genes 3' end (Fig. 9H).

The first reports of highly repetitive sequences in the genomes of both *Neisseria gonorrhoeae* and *Neisseria meningitidi* were published in 1988 (Correia et al., 1988). Nowadays, this Neisseria repeat is denoted as the DNA uptake sequence (DUS) (Duffin and Seifert, 2010). *Neisseria* spp. have been shown to preferentially take up and transform their DNA by recognizing this non-palindromic repeat (Duffin and Seifert, 2010). DUS was shown to affect transformation by limiting DNA uptake and recombination in favor of homologous DNA (Frye et al., 2013).

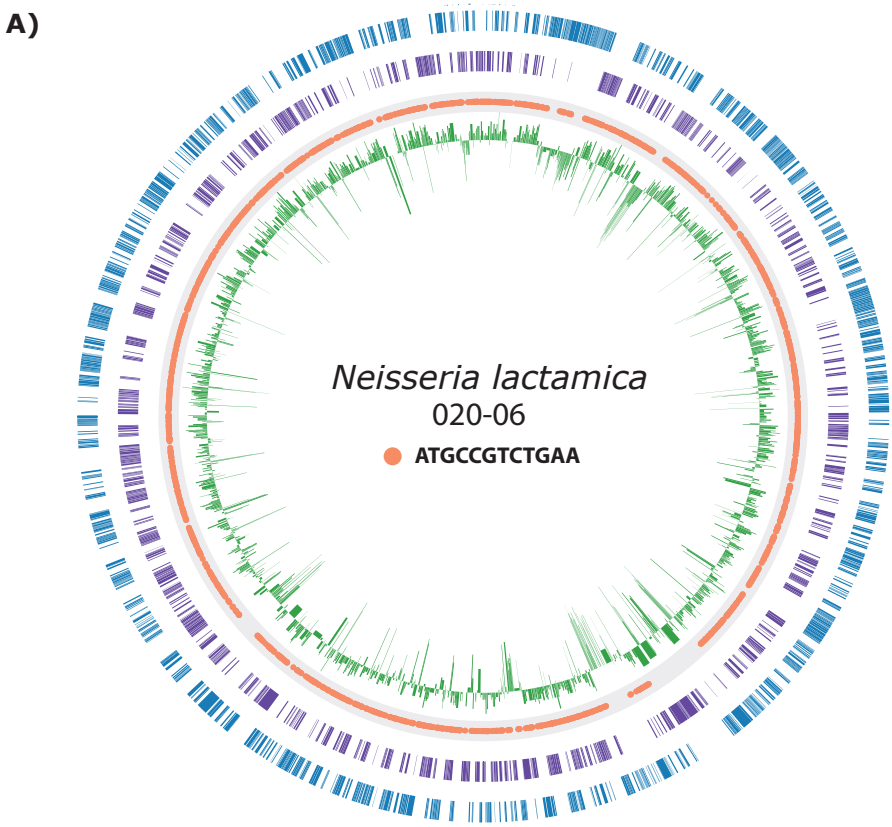


Figure 9 (A). The distribution profile of ATGCCGTCTGAA in the genome of *Neisseria lactamica* 020-06. A) Circular map of *Neisseria lactamica* 020-06 with the identified locations of the ATGCCGTCTGAA repeated sequence (orange). In addition, the following data were included: i) ORFs on the plus strand (blue), ii) ORFs on the minus strand (purple) and iii) GC-percentage (green).

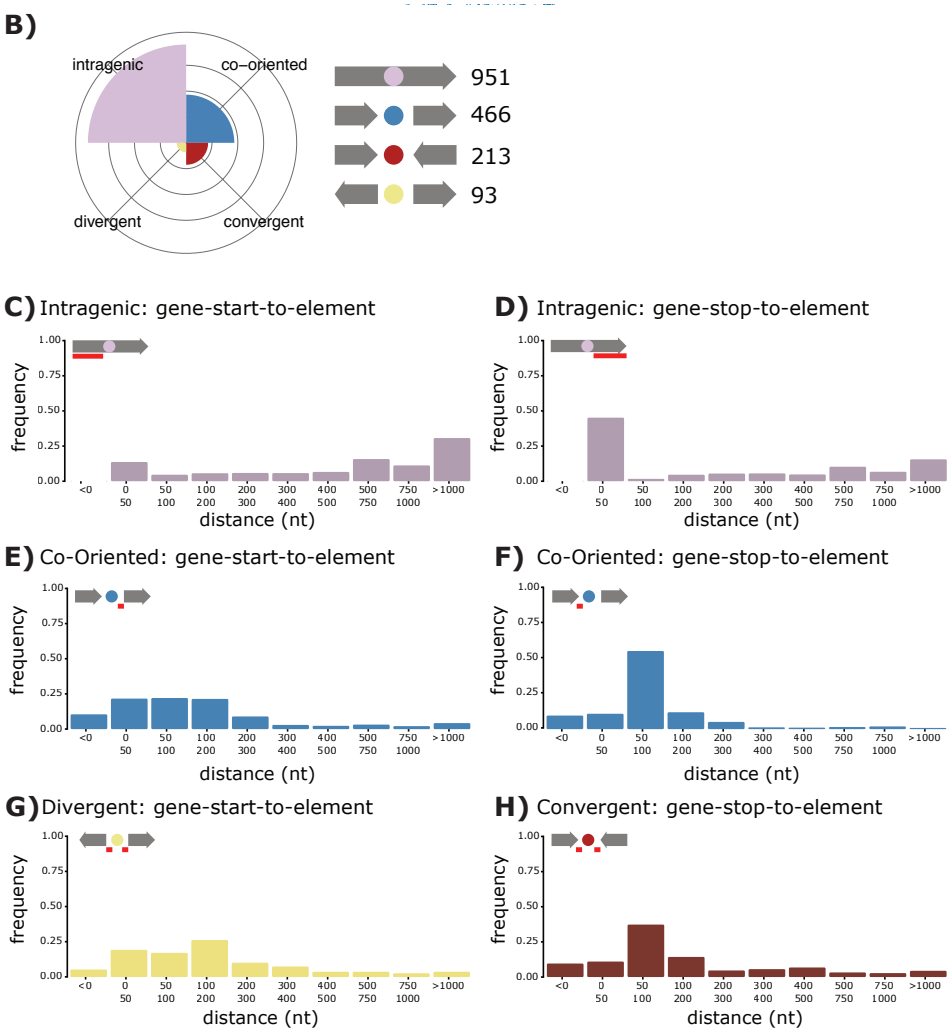


Figure 9 (B-H). The distribution profile of ATGCCGTCTGAA in the genome of *Neisseria lactamica* 020-06. **B)** Local gene organization of genes adjacent to the ATGCCGTCTGAA sequence. **C)** and **D)** Distribution of the distance of intragenic copies to the start and stop codon of the gene in which they are located. **E)** and **F)** Distribution of the distance of copies located in co-oriented intergenic regions to the downstream start codon and upstream stop codon. **G)** Distribution of the distance of copies located in divergent intergenic regions to the nearest start codon of a gene. **H)** Distribution of the distance of copies located in convergent intergenic regions to the nearest stop codon of a gene.

Highly repetitive motif (HRM) in Lactococcus lactis

The overlapping sequences (TT)CTGTCAGTAA(AA) (TTCTGTCAGTAA, TCTGTCAGTAAA and CTGTCAGTAAAA) were abundant in and highly specific for the species *Lactococcus lactis* (Table 3). The occurrence in other species, including the close relative *Lactococcus garvieae* was lower than 10 except for *Clostridium clariflavum* DSM 19732, a *Firmicute*, in which we identified 15 copies. TTCTGTCAGTAA is part of a previously described 13 bp sequence, WWNTTACTGACRR (reverse complement: YYGTCAGTAANWW), which was called highly repetitive motif (HRM) by (Mrázek et al., 2002). It was speculated that the sequence had a possible dual role: i) in transcription termination, because of its occurrence downstream of genes; and ii) as a protein-binding site, since the spacings of consecutive HRMs were supposedly consistent with the DNA helical period (Mrázek et al., 2002). However, the distribution and relative location we observed of the repeated sequence within the genome of *L. lactis* subsp. *lactis* IL1403 seemed to refute these hypotheses (Fig. 10A). The sequence was found distributed uniformly over the chromosome of *L. lactis subsp. lactis* IL1403 (Fig. 10A). The TTCTGTCAGTAA sequence was found located mostly inside coding sequences (206 copies), whereas 135 copies were found between genes in operons (Fig. 10B). About ~45 % of the intragenic copies were located closely (<50 nt) to the 3' end of the gene (Fig. 10C). Both the genome-wide distribution and the distribution with respect to the adjacent genes resembled that reported for the DNA uptake sequence (DUS) in *Neisseria* spps (Fig. 9). A function analogous to that of the DUS sequence is therefore plausible.

A different highly specific repeated sequence, ACCCGAATTGCT (reverse complement: AGCAATTCGGGT) was found 101 times in the genome of *L. lactis* subsp. *cremoris* SK11. This sequence was highly specific for the *cremoris* sub-species as it was not found in the *lactis* sub-species. 15 copies were found in the genome of *L. garvieae* ATCC 49156 (Table 3). Other characteristic abundant repeated sequences were found present within the family of Streptococcaceae. For example, AAAATCAAAGAG was present in at least 50 copies in 16 out of 58 Streptococcal genomes, whereas it was present less than 10 times in the other 42 Streptococcal genomes.

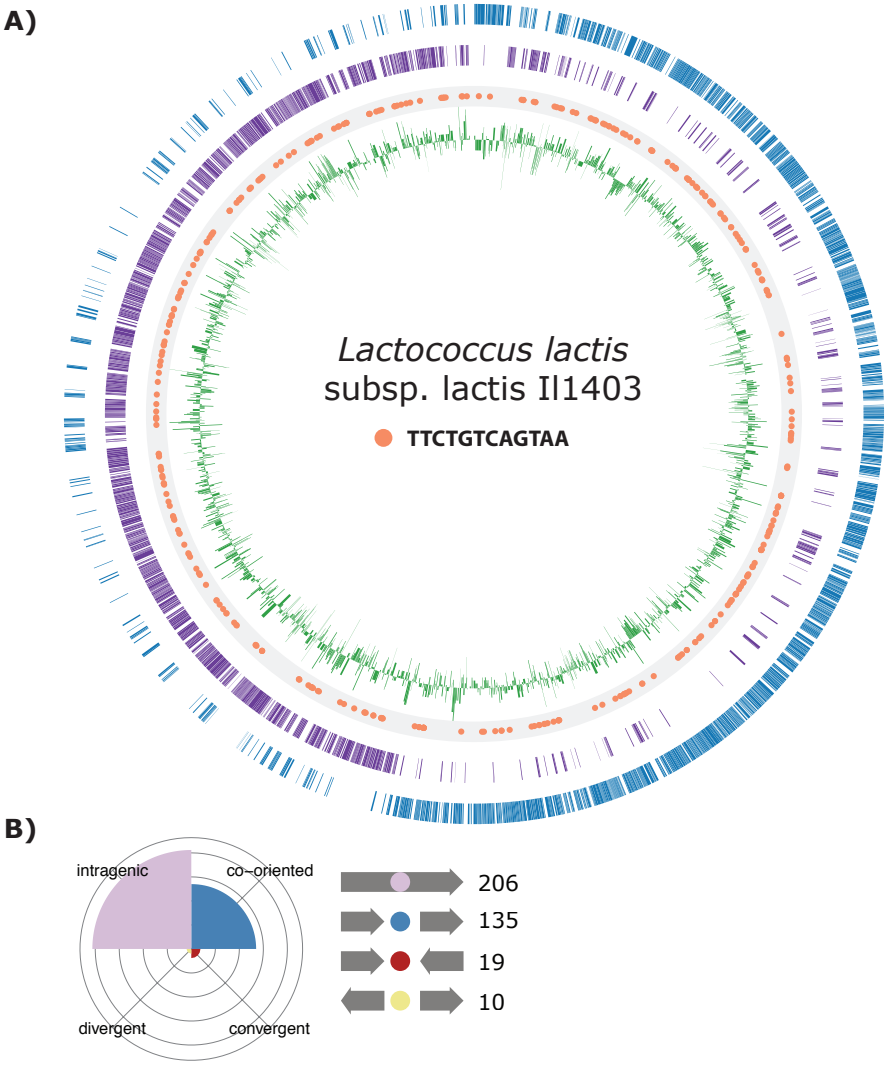


Figure 10 (A-B). The distribution profile of TTCTGTCAGTAA in the genome of *L. lactis* subsp. lactis Il1403. A) Circular map of *L. lactis* subsp. lactis Il1403 with the identified locations of the TTCTGTCAGTAA repeat (orange). In addition, the following data were included: i) ORFs on the plus strand (blue), ii) ORFs on the minus strand (purple) and iii) GC-percentage (green). **B)** Local gene organization of genes adjacent to the TTCTGTCAGTAA sequence.

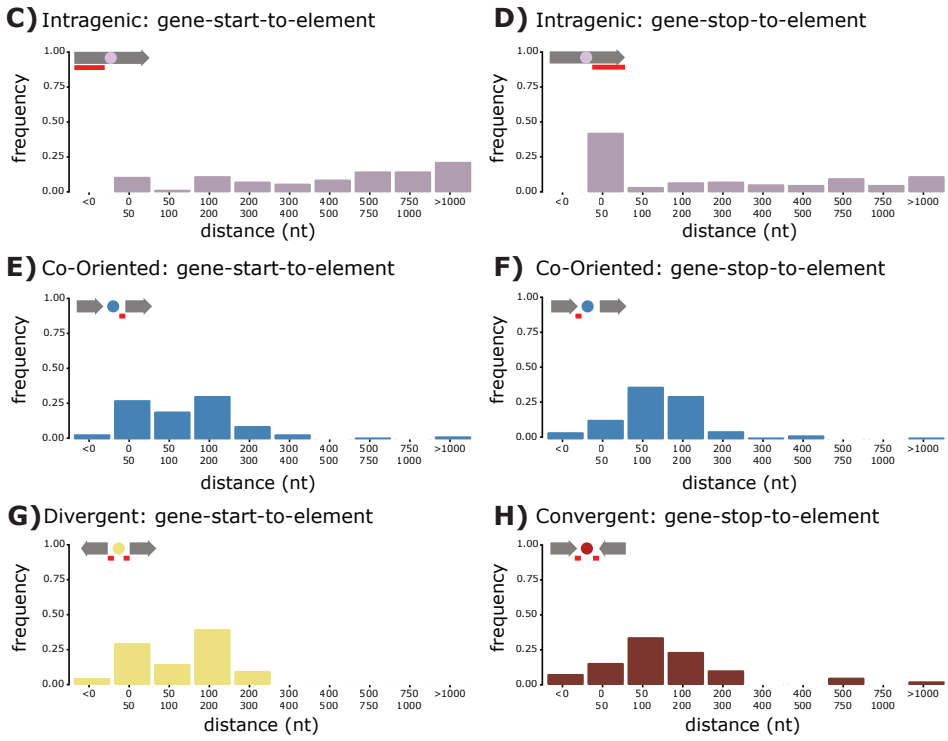


Figure 10 (C-H). The distribution profile of TTCTGTCAGTAA in the genome of *L. lactis* subsp. lactis Il1403. **C)** and **D)** Distribution of the distance of intragenic copies to the start and stop codon of the gene in which they are located. **E)** and **F)** Distribution of the distance of copies located in co-oriented intergenic regions to the downstream start codon and upstream stop codon. **G)** Distribution of the distance of copies located in divergent intergenic regions to the nearest start codon of a gene. **H)** Distribution of the distance of copies located in convergent intergenic regions to the nearest stop codon of a gene.

Table 3. Highly specific repeats within the genus of *Lactococcus lactis* and *Lactococcus garvieae* . In addition three non *Lactococcus* species with most hits have been included.

Species	TTCTGTCAGTAA	TCTGTCAGTAAA	CTGTCAGTAAAA	ACCCGAATTGCT
L. garvieae ATCC 49156	0	2	2	15
L. lactis subsp. lactis Il1403	370	344	262	2
L. lactis subsp. cremoris SK11	210	220	141	101
L. lactis subsp. cremoris MG1363	223	230	146	4
L. lactis subsp. lactis KF147	364	337	260	0
C. clariflavum DSM 19732	15	1	0	0
S. denitrificans OS217	9	2	0	1
X. nematophila ATCC 19061	4	4	8	0

Rickettsiaceae and Rhizobiales specific repeat

Various repeated sequences were found to be highly specific for some species within the families of *Rickettsiaceae* and *Bradyrhizobiaceae* (*Alphaproteobacteria*) (Table S1; https://figshare.com/s/f707d70e804d07179e99). Among these was CGTCATTGCGAG, which was (highly-) abundant in 21 *Rickettsia* species. Within these 21 genomes, the number of copies ranged from 86 (*Rickettsia akari*) to 533 (*Rickettsia felis*), with an average of 173.33. In contrast, in four of the Rickettsia genomes in our dataset (*Rickettsia prowazekii*, *Rickettsia typhi*, *Rickettsia Canadensis* str. CA410 and *Rickettsia Canadensis* str. McKiel) only 1, 1, 9 and 9 copies were found, respectively. The identified sequence relates to first half of the species-specific REP that was identified in *Rickettsia conorii*: TATGTCATTCCCGGCAAAGCGGGAATCCAGT (identical nucleotides underlined) (Tobes and Ramos, 2005). This REP-sequence was found to occur 237 times (with some variability) in *Rickettsia conorii,* for which 81% were found in the intergenic regions (Tobes and Ramos, 2005). We identified 172 copies of the sequence CGTCATTGCGAG, of which 67.4% were located in the intergenic regions. A similar distribution was observed in *Rickettsia felis*, where 70.9% of the 533 copies were found in the intergenic regions. A relatively large number of 155 copies was found located in coding sequence. The presence of this kind of 'selfish DNA' in coding sequences in *Rickettsia* species was already observed in 2000 and then denoted as *Rickettsia* palindromic elements (Ogata et al., 2000).

The repeated sequence CGTCATTGCGAG is also abundant in the family of B*radyrhizobiaceae*, where it occurred in a range from 240 copies in *Bradyrhizobium* sp. BTAi1 to only a single copy in *Nitrobacter winogradskyi*. It was also present in 29 copies in the genome of *Parvibaculum lavamentivorans*, which is part of the family of *Phyllobacteriaceae*. The 29 copies of this repeat were distributed throughout the whole chromosome (Fig. S3), whereas within the other genomes belonging to the family of *Phyllobacteriaceae* no or only a single copy was found. The latter might suggest that the occurrence of the repeat in *P. lavamentivorans* was the result of horizontal gene transfer, but the uniform occurrence throughout the chromosome makes this unlikely. Alternatively, the species *Parvibaculum lavamentivorans* might be incorrectly taxonomically classified. Another explanation would be that the use of this particular repeated sequence has evolved independently.

Conclusion

In this study we identified overrepresented dodecamers (sequences of 12 nt) within the intergenic regions of 1516 prokaryotic genomes to identify possible new marker sequences and to uncover potential new sequences with a structural or functional role. It can be challenging to characterize newly identified (repeated) sequence elements. Therefore, we formulated a strategy to create a distribution profile for every repeated sequence based on: i) the abundance and genome-wide distribution, ii) the taxonomic distribution and iii) the distribution with respect to the local gene organization.

The repeated sequences that were described and discussed were used to illustrate the application of the formulated strategy to characterize newly identified (repeated) sequence elements in terms of distribution. Assigning a detailed function to individual sequences was beyond the scope of this chapter because it requires substantial additional analyses. In the next chapter we will explore the in-depth analysis of the biological role of a particular repeated sequence, the REP sequences that are abundantly found in the genome of the model organism *Escherichia coli*.

A total of 22755 overrepresented sequences were identified, from which 583 occurred at least 40 times in one or more individual genomes. The most widely spread and abundant repeated sequences we found were Adenine-rich. As expected, the number of these A-rich sequences correlated to the GC% of the genomes. Yet, their abundance in individual genomes implied a functional role. They were found uniformly distributed in the genomes in which they were identified and, although they were found mostly within intragenic regions, they did not seem to exhibit a positional bias with respect to local gene organization. These findings indicate that the A-rich sequences might have a global, but probably not gene-organization related function (e.g. gene transcription regulation) in the prokaryotic genomes. Interestingly, the taxonomic occurrence of the abundant T-rich repeat clearly correlated to the A-rich repeats, yet their genomic distribution did not seem correlated.

Other sequences, such as the REP related sequences in *Gammaproteobacteria*, did not show a bias in genome-wide distribution, but showed a clear bias in terms in distribution with respect to local gene organization. The REP-related sequences were primarily found in the intergenic regions, but sparsely between divergent gene-pairs and abundantly between convergent gene-pairs. This bias can be related to a distinct function of the REP sequences, as they might enable simultaneous transcription of convergent gene-pairs (see next chapter for details).

In addition, we have been able to identify sequences that are highly specific for various species and strains. The repeat sequence (A)ATGCCGTCTGAA was highly abundant within the genus *Neisseria*. Nowadays, this *Neisseria* repeat is denoted as the DNA uptake sequence (DUS) (Duffin and Seifert, 2010). DUS was shown to affect transformation by limiting DNA uptake and recombination in favor of homologous DNA (Frye et al., 2013). We found an abundant species-specific sequence TTCTGTCAGTAA (also denoted as Highly repetitive motif; HRM) with the same distribution profile in the lactic acid bacterium *Lactococcus lactis,* thus pointing to a similar function. We conclude from the distribution profile that the previous hypothesis concerning the function of HRM (Mrázek et al., 2002) is probably not correct. Some *Lactococcus lactis* genomes harbored another highly abundant repeated sequence ACCCGAATTGCT. This repeated sequence was highly specific for *L. lactis* subsp. *cremoris* SK11 and therefore possible suited for strain identification.

By identifying 583 overrepresented dodecamers we found sequences with varying and sometimes intriguing abundance and distribution profiles making them potentially very useful in the identification of specific species or strains. We have defined a potential strategy to characterize the sequence element based on the examination of the distribution over the chromosome and with respect to gene organization. The strategy may serve as a foundation for further in-depth analyses of the function of the sequences. The provided distribution profile can be used to select sequences that match the characteristics of sequences for which the function is known. Vice versa, the distribution profile can be used to narrow down the list of potential functions. We have shown that the different distribution characteristics are relevant to describe different types of sequences. By a using combination of different perspectives, i.e. a broad taxonomic view, a genome-wide view and view focused on local gene organization we have been able to provide a valuable description of the identified repeated sequences.

Supporting information

Supplementary data is freely available online:
https://figshare.com/s/f707d70e804d07179e99

References

Achaz, G., Rocha, E.P.C., Netter, P., and Coissac, E. (2002). Origin and fate of repeats in bacteria. *Nucleic Acids Res.* **30,** 2987–2994.

Bachellier, S., Saurin, W., Perrin, D., Hofnung, M., and Gilson, E. (1994). Structural and functional diversity among bacterial interspersed mosaic elements (BIMEs). *Mol. Microbiol.* **12**, 61–70.

Bailey, T.L., Williams, N., Misleh, C., and Li, W.W. (2006). MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.* **34**, W369–W373.

Cho, N.-H., Kim, H.-R., Lee, J.-H., Kim, S.-Y., Kim, J., Cha, S., Kim, S.-Y., Darby, A.C., Fuxelius, H.-H., Yin, J., et al. (2007). The Orientia tsutsugamushi genome reveals massive proliferation of conjugative type IV secretion system and host–cell interaction genes. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 7981–7986.

Correia, F.F., Inouye, S., and Inouye, M. (1988). A family of small repeated elements with some transposon-like properties in the genome of *Neisseria gonorrhoeae*. *J. Biol. Chem.* **263**, 12194–12198.

Davenport, C.F., and Tümmler, B. (2010). Abundant Oligonucleotides Common to Most Bacteria. *PLoS ONE* **5**, e9841.

Delihas, N. (2007). Enterobacterial small mobile sequences carry open reading frames and are found intragenically--evolutionary implications for formation of new peptides. *Gene Regul. Syst. Biol.* **1**, 191–205.

Duffin, P.M., and Seifert, H.S. (2010). DNA uptake sequence-mediated enhancement of transformation in *Neisseria gonorrhoeae* is strain dependent. *J. Bacteriol.* **192**, 4436–4444.

Feschotte, C., and Mouchès, C. (2000). Evidence that a Family of Miniature Inverted-Repeat Transposable Elements (MITEs) from the *Arabidopsis thaliana* Genome Has Arisen from a pogo-like DNA Transposon. *Mol. Biol. Evol.* **17**, 730–737.

Filée, J., Siguier, P., and Chandler, M. (2007). Insertion Sequence Diversity in Archaea. *Microbiol. Mol. Biol. Rev.* **71**, 121–157.

Francke, C., Groot Kormelink, T., Hagemeijer, Y., Overmars, L., Sluijter, V., Moezelaar, R., and Siezen, R.J. (2011). Comparative analyses imply that the enigmatic sigma factor 54 is a central controller of the bacterial exterior. *BMC Genomics* **12**, 385.

Frye, S.A., Nilsen, M., Tønjum, T., and Ambur, O.H. (2013). Dialects of the DNA Uptake Sequence in Neisseriaceae. *PLoS Genet* **9**, e1003458.

Gevers, D., Huys, G., and Swings, J. (2001). Applicability of rep-PCR fingerprinting for identification of *Lactobacillus* species. *FEMS Microbiol. Lett.* **205**, 31–36.

Gilson, E., Perrin, D., and Hofnung, M. (1990). DNA polymerase I and a protein complex bind specifically to *E. coli* palindromic unit highly repetitive DNA: implications for bacterial chromosome organization. *Nucleic Acids Res.* **18,** 3941–3952.

Gilson, E., Saurin, W., Perrin, D., Bachellier, S., and Hofnung, M. (1991). Palindromic units are part of a new bacterial interspersed mosaic element (BIME). *Nucleic Acids Res.* **19**, 1375–1383.

Grissa, I., Vergnaud, G., and Pourcel, C. (2007). The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics* **8**, 172.

Han, K., Li, Z., Peng, R., Zhu, L., Zhou, T., Wang, L., Li, S., Zhang, X., Hu, W., Wu, Z., et al. (2013). Extraordinary expansion of a *Sorangium cellulosum* genome from an alkaline milieu. *Sci. Rep.* **3**, 2101.

Haran, T.E., and Mohanty, U. (2009). The unique structure of A-tracts and intrinsic DNA bending. *Q. Rev. Biophys.* **42**, 41–81.

Hernández-Salmerón, J.E., Valencia-Cantero, E., and Santoyo, G. (2013). Genome-wide analysis of long, exact DNA repeats in *rhizobia*. *Genes Genomics* **35**, 441–449.

Higgins, C.F., Ames, G.F.-L., Barnes, W.M., Clement, J.M., and Hofnung, M. (1982). A novel intercistronic regulatory element of prokaryotic operons. *Nature* **298**, 760–762.

Hulton, C.S., Higgins, C.F., and Sharp, P.M. (1991). ERIC sequences: a novel family of repetitive elements in the genomes of *Escherichia coli*, *Salmonella typhimurium* and other Enterobacteria. *Mol. Microbiol.* **5**, 825–834.

Jansen, R., Embden, J.D.A. van, Gaastra, W., and Schouls, L.M. (2002). Identification of genes that are associated with DNA repeats in prokaryotes. *Mol. Microbiol*. **43**, 1565–1575.

Koonin, E.V., and Wolf, Y.I. (2008). Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Res.* **36**, 6688–6719.

Koressaar, T., and Remm, M. (2012). Characterization of Species-Specific Repeats in 613 Prokaryotic Species. *DNA Res.* **19**, 219–230.

Lagesen, K., Ussery, D.W., and Wassenaar, T.M. (2010). Genome update: the 1000th genome--a cautionary tale. *Microbiol. Read. Engl.* **156**, 603–608.

Mahillon, J., and Chandler, M. (1998). Insertion Sequences. *Microbiol. Mol. Biol. Rev.* **62**, 725–774.

Makarova, K.S., Grishin, N.V., Shabalina, S.A., Wolf, Y.I., and Koonin, E.V. (2006). A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biol. Direct* **1**, 7.

Mrázek, J., Gaynon, L.H., and Karlin, S. (2002). Frequent oligonucleotide motifs in genomes of three streptococci. *Nucleic Acids Res.* **30**, 4216–4221.

Newbury, S.F., Smith, N.H., and Higgins, C.F. (1987). Differential mRNA stability controls relative gene expression within a polycistronic operon. *Cell 51*, 1131–1143.

Ogata, H., Audic, S., Barbe, V., Artiguenave, F., Fournier, P.-E., Raoult, D., and Claverie, J.-M. (2000). Selfish DNA in Protein-Coding Genes of *Rickettsia. Science* **290**, 347–350.

Overmars, L., van Hijum, S.A.F.T., Siezen, R.J., and Francke, C. (2015a). CiVi: circular genome visualization with unique features to analyze sequence elements. *Bioinforma. Oxf. Engl.* **31**, 2867–2869.

Overmars, L., Siezen, R.J., and Francke, C. (2015b). A Novel Quality Measure and Correction Procedure for the Annotation of Microbial Translation Initiation Sites. *PLoS ONE* **10**, e0133691.

Pruitt, K.D., Tatusova, T., Brown, G.R., and Maglott, D.R. (2012). NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.* **40**, D130–D135.

Rademaker, J.L.W., Louws, F.J., Versalovic, J., De Bruijn, F.J., Kowalchuk, G.A., Head, I.M., Akkermans, A.D.L., van Elsas, J.D., and others (2004). Characterization of the diversity of ecologically important microbes by rep-PCR genomic fingerprinting. *Mol. Microb. Ecol. Man.* Vol. **1** 2 611–643.

Rocha, E.P., Danchin, A., and Viari, A. (1999). Functional and evolutionary roles of long repeats in prokaryotes. *Res. Microbiol.* **150**, 725–733.

Sorek, R., Kunin, V., and Hugenholtz, P. (2008). CRISPR--a widespread system that provides acquired resistance against phages in Bacteria and Archaea. *Nat. Rev. Microbiol.* **6**, 181–186.

Swinger, K.K., and Rice, P.A. (2004). IHF and HU: flexible architects of bent DNA. *Curr. Opin. Struct. Biol.* **14**, 28–35.

Tobes, R., and Ramos, J.-L. (2005). REP code: defining bacterial identity in extragenic space. *Environ. Microbiol.* **7**, 225–228.

Tolstorukov, M.Y., Virnik, K.M., Adhya, S., and Zhurkin, V.B. (2005). A-tract clusters may facilitate DNA packaging in bacterial nucleoid. *Nucleic Acids Res.* **33**, 3907–3918.

Touchon, M., and Rocha, E.P.C. (2007). Causes of insertion sequences abundance in prokaryotic genomes. *Mol. Biol. Evol.* **24**, 969–981.

Treangen, T.J., Abraham, A.-L., Touchon, M., and Rocha, E.P.C. (2009). Genesis, effects and fates of repeats in prokaryotic genomes. *FEMS Microbiol. Rev.* **33**, 539–571.

Van Leuven, J.T., Meister, R.C., Simon, C., and McCutcheon, J.P. (2014). Sympatric Speciation in a Bacterial Endosymbiont Results in Two Genomes with the Functionality of One. *Cell* **158**, 1270–1280.

Versalovic, J., Koeuth, T., and Lupski, J.R. (1991). Distribution of repetitive DNA sequences in Eubacteria and application to fingerprinting of bacterial genomes. *Nucleic Acids Res.* **19**, 6823–6831.

Chapter 4

Repetitive Extragenic Palindromic elements have a common topological role in the reduction of transcriptional interference

Lex Overmars, Sacha A.F.T. van Hijum, Roland J. Siezen, Christof Francke

Manuscript submitted for publication

Abstract

Repetitive Extragenic Palindromic elements (REPs) are short palindromic sequences, abundantly found in enteric bacteria. These sequences are well-conserved and show a characteristic genomic distribution. Various biological roles have been proposed, however none of them has provided a common functional denominator. To find clues regarding their biological role we have analyzed the organization of REPs on the genome of the model enteric bacterium *Escherichia coli* MG1655 in depth. We confirmed that REPs are specifically located on the non-coding part of the DNA solely between convergent and co-oriented genes and found that they exhibit a clear preference for a location close to a stop-codon. We then analyzed the sequence characteristics of the genes separated by REP elements and observed that the upstream members of co-oriented gene-pairs have a significantly higher Codon Adaptation Index (CAI). The CAI of convergent gene-pairs interspersed with REPS also appeared significantly higher when compared to their counterparts without REPs. The observed increase in CAI implies a relatively high expression level of the genes associated with REPS. Indeed a significant increase in expression was found using actual (average-) gene expressions derived from a set of 466 publicly available microarrays (M3D). The same set of experiments was used to further explore the effect of the presence of REPs on transcription of the adjacent genes. The analysis showed that the average difference in expression between convergent REP-associated gene-pairs was lower compared to non-REP-associated gene-pairs for every experiment in the data set. We thus found that REPs somehow reduce the difference in expression between neighboring genes, at an expression level that is higher on average. At the same time the expression of convergent gene-pairs and of genes located downstream of highly expressed genes is impaired by the phenomenon of transcription induced DNA supercoiling. We therefore hypothesize that REPs relax the negative effects of supercoiling on the expression level by the formation of cruciform structures, thereby allowing for a higher expression of convergent gene-pairs and of genes located downstream of highly expressed genes. Moreover, the effect of REPS on expression appears important under many conditions as a high expression of both members of many REP-associated gene-pairs was found for many physiological conditions.

Introduction

Repetitive Extragenic Palindromic elements (REPs) were discovered in the DNA of *Escherichia coli* and *Salmonella typhimurium* by (Gilson et al., 1984; Higgins et al., 1982). Later, REPs were found to appear in large numbers -in the genome of *Escherichia coli* over 600 copies are present- and thereby to occupy a significant fraction of the total intergenic-space (representing about 0.5-1% of the genome sequence). Despite the overwhelming presence, so far a common biological role of REPs has remained obscure.

The occurrence of REPs appeared not limited to *Escherichia coli* and *Salmonella typhimurium* genomes but REP-like sequences were found in all Enterobacteriales, and also in more distant species, like *Pseudomonas putida* (Aranda-Olmedo et al., 2002). An automated search of 57 gamma-proteobacterial genomes resulted in the identification of 11 additional REP-containing species (Tobes and Ramos, 2005). In this search REPs adhered to the following criteria: they were (i) located in intergenic space; (ii) palindromic; (iii) between 21 and 61 base-pairs in length; and (iv) occupying more than 0.5% of the total intergenic-space.

REPs have been divided in three different types based on small variations in their sequence: Y, Z^1 and Z^2 (Bachellier et al., 1994). REPs are commonly present as repeated units, which were called Bacterial Interspersed Mosiac Elements (Gilson et al., 1991). BIMEs were grouped in three different families, according to the associated REP types: i) BIME-1, ii) BIME-2 and iii) 'atypical' (Bachellier et al., 1994). BIME-1 elements consist of repeating Y and Z^1 REPs, whereas BIME-2 elements consist of a number of repeating Y and Z^2 REPs (Bachellier et al., 1994). BIME-1s are mostly located at the 3' end of transcription units, whereas BIME-2s are frequently found between two genes belonging to the same operon (Espéli et al., 2001).

Various functional roles have been attributed to REPs. One of the earliest functional roles ascribed to REPs was to increase mRNA stability (Newbury et al., 1987). The genes *malE* and *malF*, present on the polycistronic transcript *malEFG*, were shown to be differentially expressed. This feature was attributed to the presence of a REP element between *malE* and *malF*, which supposedly stabilized the 3'-part of the transcript (Newbury et al., 1987). In another study, several REP-containing mRNA transcripts were shown to be less prone to degradation by the RNA degradosome *in vivo* (Khemici and Carpousis, 2004). Earlier, Yang and Ames recognized that mRNA stabilization should be a consequence of the palindromic nature of REPs (Yang and Ames, 1988). Hence this should not necessarily represent the REPs primary function.

Several proteins were shown to interact with specific REP types. For instance, DNA polymerase I was found to bind to the repetitive REPs (BIME-2) located in the intergenic region of the *malB* locus *in vitro* (Gilson et al., 1990). Similarly, DNA gyrase was found to bind to repetitive REPs (BIME-2), both *in vitro* (Yang and Ames, 1988) and *in vivo* (Espéli and Boccard, 1997). Interestingly, the binding was stimulated by the HU protein (Yang and Ames, 1988). BIMEs were also proposed to facilitate Rho-dependent transcription termination (Espéli et al., 2001). In addition, a short conserved sequence separating individual REPs within a BIME-1 element was found to be recognized by integration host factor (IHF) (Boccard and Prentki, 1993; Oppenheim et al., 1993). Oppenheim and colleagues termed these specific sequences RIP-elements (repetitive IHF-binding palindromic), whereas Boccard and colleagues denoted these elements as RIB (reiterative IHF BIME). It was estimated that a significant number of these elements, 70 or 100, respectively, are present in the *E. coli* genome. Oppenheim and colleagues postulated that the inverted repeat structures present in REPs and in the RIP elements could form cruciform structures and that the binding by IHF might prevent their formation (Oppenheim et al., 1993), similar to the HU protein which was shown to interfere with the formation of particular cruciform structures by binding to an intermediate in cruciform formation (Pontiggia et al., 1993).

REPs have also been linked to recombination events and transposition (Nunvar et al., 2010; Tobes and Pareja, 2006). All 14 *E. coli* K12 IS1397 integration sites correspond to a BIME sequence (Bachellier et al., 1997). In 6 out of 19 analyzed bacterial genomes in which REPs have been found, an association was found with insertion sites of mobile elements, and it was therefore suggested that REPs are hot spots for transposition (Tobes and Pareja, 2006). In addition, the introduction of a REP within 15 nt of a termination codon was shown to cause a decrease in translation (Liang et al., 2015). A longer spacing (>15nt) caused no effect, suggesting that REPs close to the stop codon could stall ribosome movement.

So far, none of the proposed functions supply a general and uniform explanation for the biological role of REPs. Yet, the remarkably high abundance, and the strong sequence conservation and similarity between the various types, suggest a common role. A puzzling feature of REPs is their genomic distribution. Though REPs occur throughout the chromosome, REPs were solely found between convergent and co-oriented gene-pairs (Aranda-Olmedo et al., 2002; Tobes and Ramos, 2005). This observation made us realize a common role of REPs should be linked to this particularly biased distribution. We retrieved the position of all REPs on the *E. coli* genome and analyzed their distribution and gene-association. We also performed a

statistical analysis on a large set of microarray data to investigate possible large-scale effects of REPs on the expression of transcripts. We found a specific link of REPs to the expression level of the associated genes. REPs appeared to enable a relatively high expression and simultaneously a relatively small difference in expression.

Materials and methods

Genome sequence, annotation and microarray data.

The complete genome sequence and gene annotations of *Escherichia coli* K12 MG1655 were obtained from the NCBI GenBank database (ftp.ncbi.nih.gov/genomes/Bacteria/). The distribution of REPs throughout the chromosome of *E. coli* K12 MG1655 was analyzed and visualized using CiVi (Overmars et al., 2015).

REP positions and BIME classification

REP positions were derived from EcoGene (Zhou and Rudd, 2012), as available on the corresponding topic page (http://www.ecogene.org/?q=topic/172). Repeated REPs were grouped into the three different types of BIMEs as used by Bachellier and colleagues: BIME-1, BIME-2 or 'atypical' (Bachellier et al., 1994). The location of potential REPs in the *E. coli* genome was also established by performing a similar motif search as described in (Francke et al., 2011).

Association of REPs to gene-pairs

REPs were grouped according to the relative directionality of the adjacent pair of genes in to: (i) same strand noted as 'Co-oriented' (i.e. → REP →); (ii) divergent noted as 'Divergon' (i.e. ← REP →); and (iii) convergent noted as 'Convergon' (i.e. → REP ←); or (iv) inside a gene noted as 'coding region'. The maximal mutual distance between the two flanking genes was set to 350 nt. Gene-pairs exceeding this distance limit were not included in the analysis.

Association of Codon Adaptation Index to REPs

The mRNA sequences of highly expressed genes are usually biased towards codons that are recognized by the most abundant tRNA molecules. A numerical measure of this bias is the codon adaptation index (CAI) (Sharp and Li, 1987). Codon Adaptation Index (CAI) values for all *E. coli* K12 MG1655 genes were extracted from the highly expressed genes database (HEG-DB) (Puigbò et al., 2008). A Mann-Whitney *U* test was applied to test the differences between

REP containing (i.e. → REP ←) and REP-lacking (i.e. → ←) CAI distributions for convergent gene-pairs. The CAI association for REPs separating co-oriented gene-pairs was analyzed by comparing the CAI values of the gene upstream from a REP to the gene downstream from a REP. In addition, a chi-squared test was applied to compare the distribution of CAI values of REP associated genes with the complete genome of *E. coli* K12 MG1655. To apply this test 3 different definitions of a 'high' CAI value were tested: i) > 0.5, ii) > 0.6 and iii) > 0.7.

Association of experimental conditions to REP related gene expression

Normalized expression values of 466 microarray experiments in *E. coli* were obtained from the Many Microbe Microarrays database (M^{3D}; http://m3d.bu.edu) (Faith et al., 2008). The Many Microbe Microarrays database consists of microarray data from many different experiments performed under various conditions. The experiments have been categorized in seven physiological conditions by (Ma et al., 2013b), namely: exponential growth, stationary growth, anaerobiosis, heat shock, oxidative stress, nitrogen limitation and SOS response. The normalized and annotated microarray data were used to examine potential correlations between gene expression and specific conditions with REPS.

The expression ratio within each of the 466 experiments was calculated for all convergent gene-pairs by dividing the expression of the first gene by that of the second gene of the pair (locus tag order) and the absolute distance from an expression ratio of 1 was calculated (i.e. a similar expression value for the two genes in a gene-pair results in a ratio of 1 and a distance of 0). An average distance-from-ratio-one was determined for both REP-associated (i.e. → REP ←) and non REP-associated (i.e. → ←) gene-pairs for each individual microarray experiment, resulting in 466 average distance-from-ratio-one values for both groups of gene-pairs.

For each REP-associated gene a relative expression was calculated per experiment using: relative expression = (expression value in specific experiment - minimum expression value in all experiments) / (maximum expression value – minimum expression value in all experiments). REP-associated gene-pairs (AB) with a high expression were defined using the following criteria: relative expression *gene A* > 50%; relative expression *gene B* > 50% and; summed relative expression > 120%. The number of occurrences of high expression was counted for each REP-associated gene-pair and the number of highly expressed REP-associated gene-pairs was counted for each experiment.

Results

Genomic distribution of REPs in E. coli K12 MG1655

We first encountered REPs during the study of the diversity in the intergenic regions of the colanic acid biosynthesis cluster. The extracellular polysaccharide colanic acid is required for the development of *E. coli* K12 biofilm (Danese et al., 2000) and the responsible gene-cluster comprises 20 genes (Fig S1; (Stevenson et al., 1996)). The REPs seemed to separate this large gene cluster in three distinct functional gene-groups, namely the genes involved in the transport of colanic acid and other polysaccharides over the outer membrane (*wza, wzb, wzc*), and two clusters involved in the production of colanic acid (*wcaABCDEF, gmd, wcaGHI, cpsBG, wcaJ, wzxC*) and (*wcaKLM*). We decided to analyze the role of the REPs in more detail.

To study the genome-wide distribution of REPs we obtained their positions in the *E. coli* K12 MG1655 genome in two ways: (i) by using the REP sequences found in the colanic acid biosynthesis cluster and performing a genome-wide search for similar sequences (see methods), and (ii) by collecting them from the EcoGene database (Zhou and Rudd, 2012). The latter collection included the first completely and was slightly larger (590 vs 698 members). By plotting the positions on the *E. coli* chromosome we visually confirmed that REPs were uniformly distributed (Fig. 1A). Moreover, we did not observe any bias in the distribution of single REP units, BIMEs flanking an IHF binding site and BIMEs without such a binding site (Fig. 1A). However, the density of BIMEs was slightly lower within the termination of replication region. As was observed before the majority of REPs was found located within the intergenic space in between co-oriented (150 BIMEs or single REPs) and convergent (203 BIMEs or single REPs) genes (Fig. 1B). 1 In contrast, only two REPs were found in between a divergent gene-pair. We did not observe any bias in the distribution of convergent and co-oriented REPs (Fig. 1A), neither did we observe a bias in the presence of IHF binding sites within convergent or co-oriented REPs.

Most REP-containing intergenic regions consisted of one or two REPs and the maximum number of REPs in a single intergenic region was 12 (Fig. 2A). There was no distinct difference in the distribution of the number of REPs associated with convergent and co-oriented gene pairs. The size of the intergenic region only slightly correlated with the number of REPs within an intergenic region (Fig. 2B). Conceivably, the regions with higher number of REPs were larger to accommodate the repeats. Yet, the intergenic regions containing one or two elements were not always smaller. We also analyzed

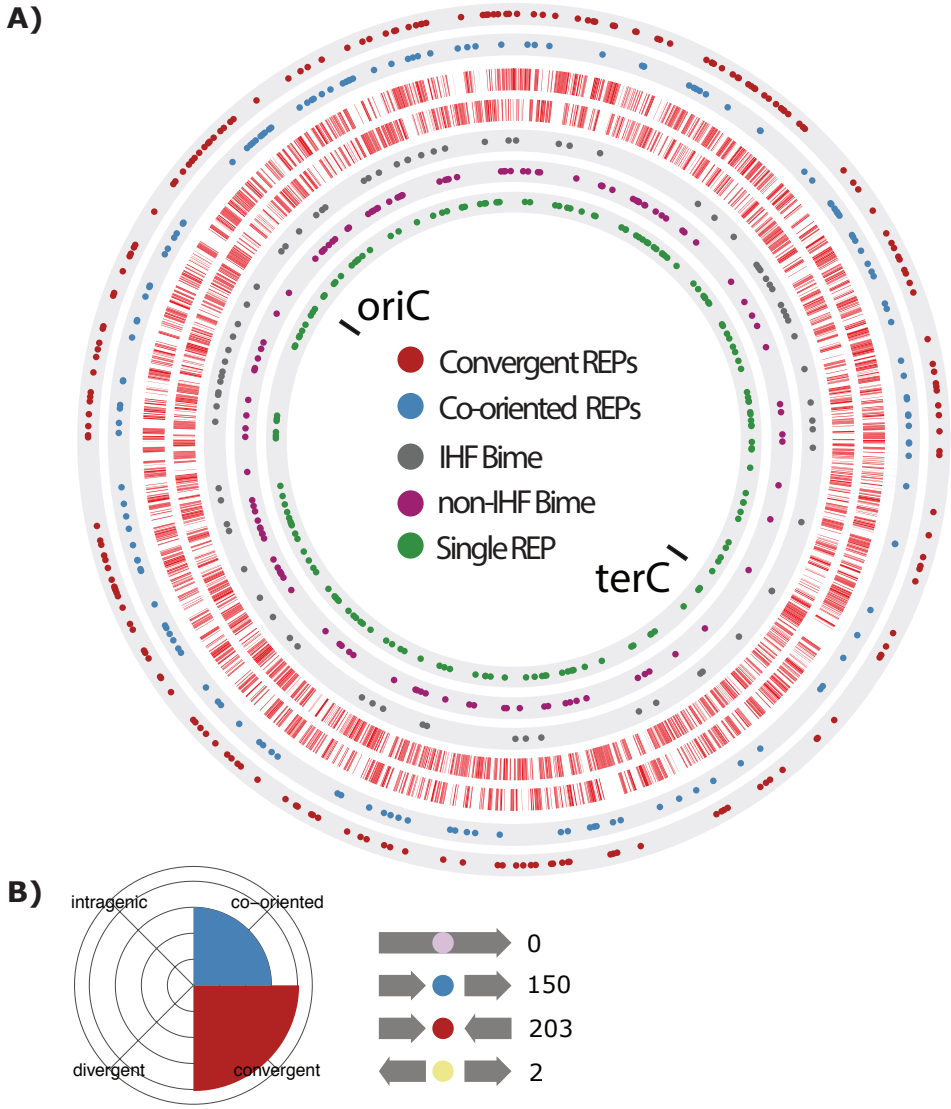


Figure 1. The global distribution of REP elements on the chromosome of *Escherichia coli* K-12 MG1655. A) Circular map of *E. coli* str. K-12 MG1655 with the EcoGene REP locations. Within the two outer rings the position of BIMEs and individual REPs elements have been colored based on the orientation of the flanking genes; convergent (dark-red) and co-oriented (blue). In addition, the ORFs on the plus (outer red) and ORFs on the minus strand (inner red) were included. The positions of: i) BIMEs with IHF binding site (grey); ii) BIMEs without binding site (purple); and iii) single element REPs (green) are represented in the inner three rings. Finally, the positions of the oriC and terC genes were indicated. **B)** Radar plot visualizing the local gene organization of genes adjacent to the REP elements.

the distances of the REPs with respect to the flanking genes (Fig. 2C-E). The distance of the REPs to the start-codon of the downstream members in the co-oriented pairs appeared highly variable (Fig. 2C). Interestingly, both the REPs located in between co-oriented gene-pairs and the REPs located in between convergent gene-pairs were located closely to the stop-codon. In case of the REPs associated to co-oriented gene-pairs >75% of REPs was located within 50 nt of the stop-codon of the upstream coding sequence (Fig. 2D), and similarly in case of REPs associated to convergent gene-pairs ~75% was located within 50 nt of one of the two stop-codons (Fig. 2E).

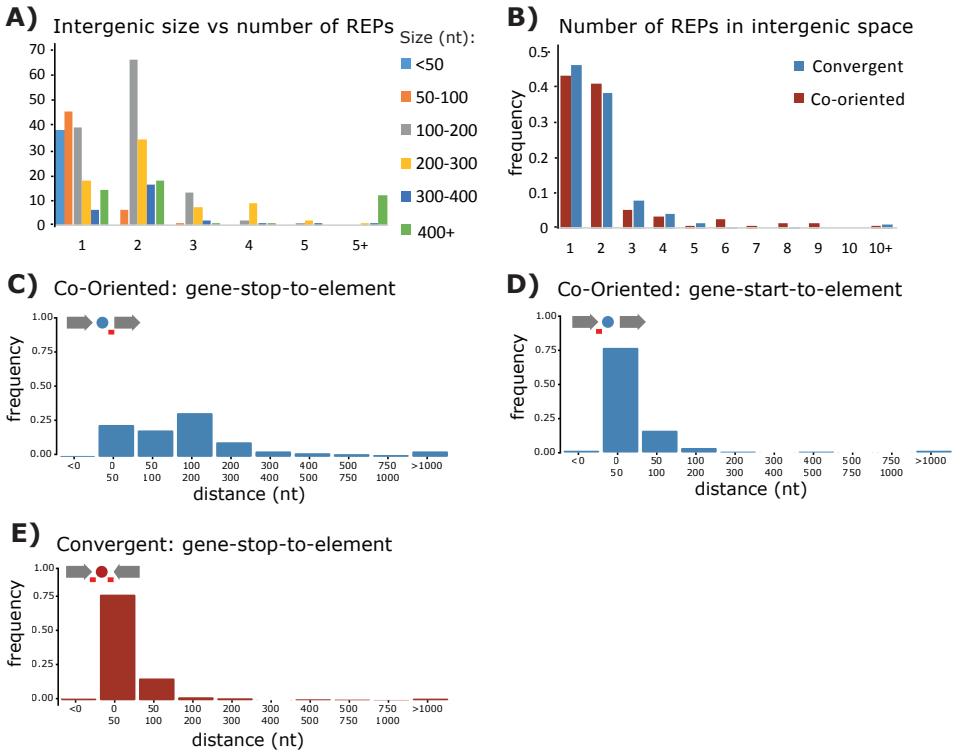


Figure 2. Distribution of REPs with respect to intergenic size and distance to adjacent genes. A) Histogram comparing the number of REP elements found in convergent- and co-oriented intergenic regions. **B)** Histrogram showing the relation between the size of intergenic regions and the number of REP elements found. **C)** and **D)** Distribution of the distance of copies located in co-oriented intergenic space to the downstream start codon and the upstream stop codon. **E)** Distribution of the distance of copies located in divergent intergenic space to the nearest start codon of a gene.

REPs are associated with a high CAI and high expression values

The codon adaptation index (CAI) is often used as an indicator of gene expression (Sharp and Li, 1987). Another family of intergenic repeats found in *E. coli*, namely: the Enterobacterial Repetitive Intergenic sequences (ERICs), was found to be associated with high CAI values in *E. coli* K12 MG1655 (Wilson and Sharp, 2006). It was reported that 10% of the genes flanking an ERIC sequence have CAI values of at least 0.7 (compared to less than 1% in the genome as a whole) whereas 21% of the adjacent genes were found to have a CAI value > 0.5 (versus 7% in the whole genome) (Wilson and Sharp, 2006). We analyzed the CAI values of the genes flanking the REPs to find out whether a comparable association could be found. CAI values for all *E. coli* K12 MG1655 genes were extracted from the highly expressed genes database (HEG-DB) (Puigbò et al., 2008). In the case of REP associated co-oriented genes we found that 20% (genes upstream of REP) and 8% (genes downstream of REP), respectively, had a CAI value of 0.7 or higher (complete genome in HEG-DB: 7%), whereas 88% (REP located at 3' end) and 69% (REP located at 5' start) of the REP flanking genes had a CAI value of 0.5 or higher (complete genome in HEG-DB: 62.7%). The CAI value of the upstream members of REP associated co-oriented gene-pairs was thus significantly higher than their downstream counterparts (p-value: 6.3e-09; Fig. 3A and Table 1). We observed the same phenomenon when comparing actual gene expression values. We calculated the average absolute expression for each gene over the 466 experiments present in the Many Microbe Microarrays database (M^{3D}; http://m3d.bu.edu) (Faith et al., 2008) and found that the upstream genes had on average significantly higher values when compared to their downstream counterparts (p- value 4.5e-05; Fig. 3B and Table 1). The effect was absent when comparing upstream versus downstream co-oriented gene-pairs without a REP (Fig. 3B).

In the case of convergent gene-pairs we compared the CAI values of the REP associated gene-pairs with those of non-REP associated gene-pairs. The summed CAI value (i.e. the sum of the two CAI values per pair) was significantly higher for REP associated convergent gene-pairs than for non-REP convergent gene-pairs (p-value 3.4e-16; Fig. 3B and Table 1). The distribution of REP associated convergent CAI values was also higher when comparing either the CAI minimum value (i.e. the lowest CAI value per pair; p-value 6.0e-10) or maximum CAI values (i.e. the highest CAI value per pair; p-value 2.2e-16). Out of 194 REP associated convergent gene-pairs, 100 had a summed CAI over 1.1. In the case of non-REP convergent gene-pairs, this was true for 94 out of 283 gene-pairs, which was significantly less (p-value 1.0e-10; fisher exact).

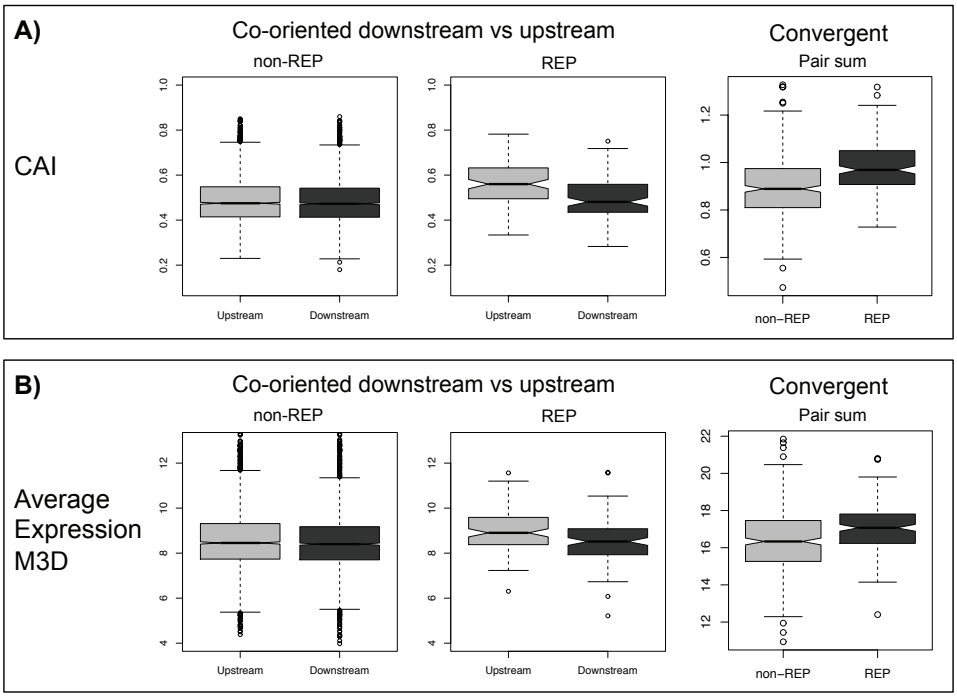


Figure 3. Expression related characteristics of genes disjointed by REPs. A) Codon adaptation index (CAI) values. In the case of co-oriented gene-pairs upstream and downstream member of the gene-pair were compared. For REP gene-pairs the upstream CAI population is significant greater, p-value 6.342e-09. In the case of convergent gene-pairs we directly compared the summed CAI values of convergent gene-pairs without REP with the convergent gene-pairs with REP. The population of summed CAI values of REP gene-pairs is significant greater, p-value 3.418e-16. **B)** Average microarray expression values. The set-up of the comparisons is similar to the CAI comparisons. In correspondence with the CAI comparison, the upstream average expression values are significant greater compared to their downstream neighbor in co-oriented REP gene-pairs, p-value 4.568e-05.

Table 1. Mann-Whitney *U* test results of the comparison of sequence- and expression- characteristics of REP associated gene-pairs.

Gene-pairs of interest		Data	P-value
Co-oriented	Upstream vs. downstream	CAI	6.3e-09
Co-oriented	Upstream vs. downstream	avg Expression	4.5e-05
Co-oriented	Upstream vs. downstream	GC percentage	2.2e-03
Convergent	REP vs. non-REP	summed CAI	3.4e-16
Convergent	REP vs. non-REP	max CAI	2.2e-16
Convergent	REP vs. non-REP	min CAI	6.0e-10

Association of physiological conditions to REP gene expression

We have analyzed the gene expression data from the M3D database to determine which aspect the expression of REP associated gene-pairs was characteristically different from the expression of non-REP associated gene pairs. First we selected the convergent gene-pairs and determined the expression ratios for each pair in all experiments and converted these ratios to (absolute-) distances to a ratio of one (see material and methods). The values were averaged for the REP associated convergent gene-pairs and the non-REP associated convergent gene-pairs within each individual experiment present in the microarray dataset. We found a lower distance for the REP associated pairs in all experiments (Fig. 4). We thus found that the presence of REPs correlated with a reduced difference in expression between neighboring genes, but at the same time also correlated with an expression level that was higher on average (Fig. 3).

We also examined whether specific physiological conditions could be linked to the expression of REP associated gene-pairs by counting the REP-associated gene-pairs with a high expression (as in that case the role of the REPs should be most prominent; see discussion). The normalized and annotated M3D microarray data was used to examine potential correlations between REP-associated gene expression and specific conditions (Table S1; https://figshare.com/s/733d34dea145f8304003). The microarray data were categorized over seven physiological conditions by (Ma et al., 2013b), namely: exponential growth, stationary growth, anaerobiosis, heat shock, oxidative stress, nitrogen limitation and SOS response. For every experiment we counted all REP-associated gene-pairs with a high expression (co-oriented and convergent; expression for both genes >50% of maximum and together >120%, see methods). We found that the number of REP-associated gene-pairs with high expression varied per experiment, but that the distribution was almost normal (Fig. 5A). Per experiment on average 45 REP associated gene-pairs showed a high relative expression. A clear difference existed between the gene-pairs, whereas some were found highly expressed in only a limited number of experiments, others were found highly expressed in the majority of experiments (Fig. 5B). Likewise, the number of highly expressed REP-associated gene-pairs varied between the physiological conditions but not in a way that linked the role of REPs to a particular condition (Fig. 5C).

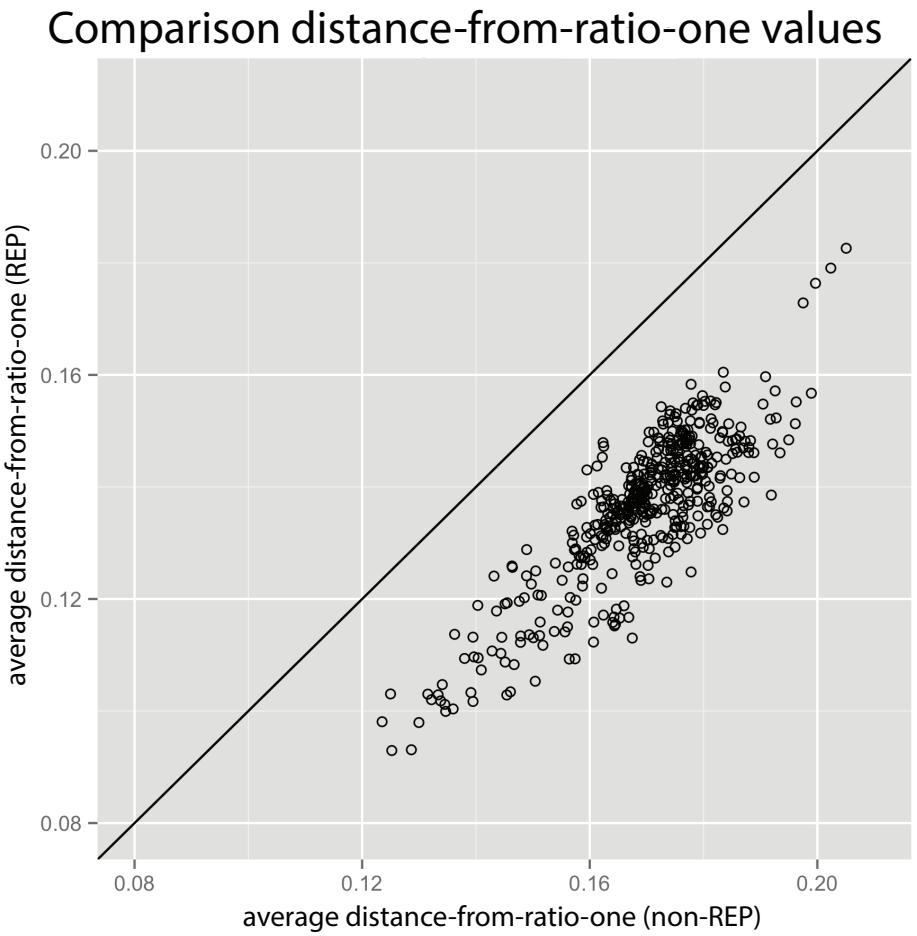


Figure 4. Comparison of the average distance-from-ratio-one values of REP versus non-REP convergent gene-pairs. Each point in this scatterplot represents an individual microarray experiment. The two averages are determined by averaging the complete set of REP and non-REP distance-from-ratio-one values, i.e. a 0 represents an identical gene expression for both members within a convergent gene-pair. The distinct difference remains the same for all experiments when we deleted the outliers with the largest differences in the non-REP convergent set of gene-pairs.

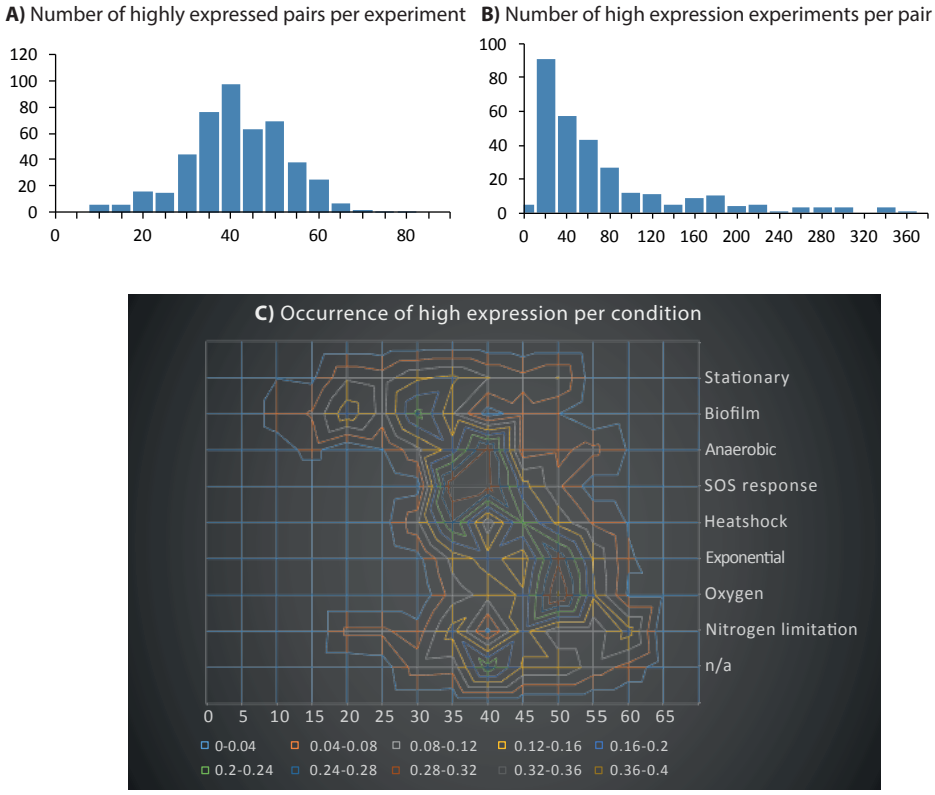


Figure 5. Occurrence of highly expressed REP-pairs. **A)** Distribution of highly expressed REP gene-pairs per experiment. **B)** Distribution of number of experiments in which an individual REP associated gene-pairs are highly expressed. **C)** Contour plot representing the fraction of highly expressed REP gene-pairs per experiment within different experimental conditions. The corresponding data can be found in Table S1A, S1B and S1F.

Four REP-associated co-oriented gene-pairs were relatively highly expressed in > 2/3 of the experiments, suggesting that expression of these genes is important in a large number of conditions (Table 2). The most frequently expressed gene-pair consisted of a transcription regulator and an Ethanolamine utilization protein, whereas the second most frequently expressed gene-pair encodes a DNA translocase and a lipoprotein carrier protein. At the same time the high expression of various REP associated gene-pairs appeared linked to particular conditions and/or clearly linked to each other. For instance, the *malEF* pair showed a high expression under varying conditions but in the majority of cases together with the *malBM* pair. This indicates that both pairs are functionally connected (which was known) -their gene-products are involved in maltose uptake and metabolism- but also that all four genes need to be highly expressed simultaneously under the given conditions. Another example was the *wcaAwcz* gene-pair, part of the colanic acid biosynthesis cluster mentioned earlier. The microarray data analysis indicated a clear connection of a high expression of the gene pair with the biofilm and stationary phase condition, though not exclusively. And showed that a high expression of the pairs *wcaKwzxC* and *yegJyegK* was in all cases associated to a high expression of the *wcaAwcz* gene-pair. The former pair is part of the colanic acid biosynthesis cluster and the function has been characterized. However, the function of the *yegJyegK* has not been elucidated so far. The data indicate that the gene pair somehow must be related to colanic acid production and export in *E. coli* K12 MG1655.

Table 2. REP associated gene-pairs that show a correlated expression[#]. A) Frequently highly expressed REP-associated gene-pairs and their products; **B)** gene-pairs that have a correlated pattern of high expression experiments with the *malEF* gene-pair; and **C)** gene-pairs that have a correlated pattern of high expression experiments with the colanic acid related gene-pair b2060_b2059.

Orientation	Pair	Locus	Product	Gene	# Exp
A) High expression in > 310 experiments					
co-ooriented	b2438_b2437	b2437	HTH-type transcriptional regulator	eutR	356
		b2438	Ethanolamine utilization protein	eutK	356
co-oriented	b0890_b0891	b0890	DNA translocase	ftsK	335
		b0891	Outer-membrane lipoprotein carrier protein (P20)	lolA	335
co-oriented	b0684_b4637	b0684	Flavodoxin-1	fldA	333
		b4637	Fur leader peptide	uof	333
co-oriented	b0027_b0028	b0027	Lipoprotein signal peptidase (EC 3.4.23.36)	lspA	322
		b0028	PPIase (EC 5.2.1.8)	fkpB	322
B) Correlated with b4034_b4033					
co-ooriented	b4034_b4033	b4033	Maltose transport system permease protein	malF	69
		b4034	Maltose-binding periplasmic protein (MBP)	malE	
co-ooriented	b4036_b4037	b4036	Maltoporin	malB	64(75)
		b4037	Maltose operon periplasmic protein	malM	
C) Correlated with b2060_b2059					
co-ooriented	b2060_b2059	b2059	Putative colanic acid biosynthesis glycosyl transferase	wcaA	82
		b2060	Tyrosine-protein kinase (EC 2.7.10.-)	wzc	
convergent	b2072_b2071	b2071	Uncharacterized protein	yegJ	19(20)
		b2072	Uncharacterized protein	yegK	
co-ooriented	b2046_b2045	b2045	Colanic acid biosynthesis protein	wcaK	19(22)
		b2046	Lipopolysaccharide biosynthesis protein	wzxC	

The analysed data can be found in Table S1C and Table S1D.

Discussion and Conclusion

In this study we aimed to uncover the primary role of REPs by analyzing their positional distribution and analyzing the sequence characteristics and expression dynamics of the neighboring genes. *E. coli* REPs are uniformly distributed over the genome (Fig. 1A) and have a clear positional preference with respect to the orientation of adjacent genes. Both observations were established before (Aranda-Olmedo et al., 2002; Bachellier et al., 1999). They are found exclusively between co-oriented and convergent gene-pairs suggestive of a role that relates to a phenomenon that is specific for these two orientations (Fig. 1B). We found the REPs associated with convergent genes are mostly located closely to the 3' end of one the adjacent genes (Fig. 2D-E), REPs located in between co-oriented pairs of genes are mostly found close to the 3' end of the upstream gene. The particular preference for a position near the stop hints at a functional connection of the REPs to a process that includes the gene 'upstream'. In addition, we established a clear association between REP elements and highly expressed genes. This was both reflected in the CAI values of REP associated genes as well as in the expression values in a large compendium of microarray experiments. More specifically, the upstream members of REP associated co-oriented genes had a higher expression value and CAI, whereas the genes in REP associated convergent gene-pairs have higher expression when compared to non-REP convergent gene-pairs. The above observations in which REPs are solely linked to co-oriented and convergent gene-pairs with high CAI values and high expression suggest a connection to the process of transcription. Moreover, the palindromic nature of REPs implies a relation to structure. In fact, the process of transcription generates forces that have an impact on local DNA structure via effects on DNA supercoiling.

The chromosome of *Escherichia coli* is maintained in a negatively supercoiled state and under a tight homeostatic control (Snoep et al., 2002). This is achieved by DNA topoisomerase enzymes that modulate the supercoiling. In *E. coli*, the level of supercoiling is primarily maintained by DNA gyrase and topoisomerases I. DNA gyrase introduces negative supercoils and removes positive supercoils (Menzel and Gellert, 1983), while topoisomerase I removes negative supercoils (Wang, 1985). Nevertheless, the supercoiling state can be affected by various environmental changes, such as osmotic stress, oxygen tension, temperature changes and nutritional shifts (Rui and Tse-Dinh, 2003). For example, the overall level of negative superhelicity decreases from exponential to stationary growth phase (Balke and Gralla, 1987). *E. coli* stationary-phase cells contain relaxed DNA molecules and recover their DNA negative supercoiling state using DNA gyrase once

nutrients become available (Gutiérrez-Estrada et al., 2014). Translocating proteins, such as RNA polymerase can affect local DNA supercoiling levels. Liu & Wang described the phenomenon of transcription driven DNA supercoiling more than two decades ago (Liu and Wang, 1987). Within this model, the transcribing polymerase acts as a torsional motor that generates positive superhelical stress ahead of the transcriptional bubble and negative stress behind it. Indeed, the presence of transcription induced DNA supercoiling was demonstrated for both prokaryotes (Tsao et al., 1989) and eukaryotes *in vitro* (Dröge, 1993; Ostrander et al., 1990). It was shown that transcription is affecting DNA structure on a large-scale *in vivo* (Krasilnikov et al., 1999). Genome-wide measurements of the supercoiling level demonstrated that gene clusters of several kilobases experience negative supercoiling correlated to the transcription level (Teves and Henikoff, 2014). The presence of the torsion-mediated interaction between the transcription of neighboring genes is often referred to as 'transcriptional interference'. An *in vivo* study reported that a resisting torque indeed slowed RNA polymerase and increased its pause frequency and duration in *E. coli* (Ma et al., 2013a). A quantitative model on transcription induced supercoiling showed that divergent promoters favor the expression of their (divergent-) neighbor, whereas convergent promoters are mutually repressive (Meyer and Beslon, 2014). The torsional coupling proposed in the model implies that any two genes with a distance less than ~3000 bases experience a mutual influence. The authors suggest that torsional coupling plays an important role in genetic regulation, and might favor the orientation-dependent co-localization of genes involved in similar functions, which need to be expressed together (Meyer and Beslon, 2014). The latter is supported by the observation that the preservation of this gene-regulatory capacity of supercoiling seems to be a driving force in the evolution of chromosomal gene order (Sobetzko, 2016).

REPs are highly overrepresented within convergent intergenic regions, separating gene-pairs that are thus mutually repressive in the model of Beyer and colleagues (Meyer and Beslon, 2014). Moreover, because of their repressive- co-localization these genes might be less likely to be functionally similar or linked. However, one could argue that there might be conditions in which the (conflicting-) expression of both genes is required for an adequate response to environmental changes. If REPs are indeed linked to transcription induced supercoiling as our observations suggest, such responses could very well justify the presence of REPs.

Transcription driven negative supercoiling can induce the formation of cruciform DNA (Dayn et al., 1992). Cruciforms were observed in cells undergoing various stresses that increase torsional tension *in vivo*, such as

inhibition of protein synthesis, anaerbiosis, and osmotic shock (Dayn et al., 1991). The formation of cruciforms can in turn reduce negative supercoiling (Krasilnikov et al., 1999). Another study already found cruciform motifs to be strongly enriched in intergenic regions separating convergent gene-pairs (Du et al., 2013). Du and colleagues predicted the position of cruciform motifs using the tool 'Inverted Repeat Finder' (threshold scores exceeding 16 and loop length between 1 and 10 bp). Almost all positions of the REPs associated with convergent gene-pairs were present in the resulting list of identified cruciform motifs. In fact, REP_{325} was shown to generate cruciform structures in a supercoiled state *in vitro* (Qian et al., 2015). The potential of REP elements to form cruciform structures means that, by their secondary structure, they could modify the local coiling state of the DNA. Under normal conditions the chromosome of *E. coli* is negatively supercoiled (Miller† and Simons, 1993). However, the overall level of supercoiling is dynamic and can vary under different conditions. Because negative supercoiling drives the formation of cruciform structures, cruciform structures could be the default state of REPs. This in turn means that, when positive supercoiling arises in front of the transcription complex during transcription of a (convergent-) gene, the cruciform structure could become unfolded, thereby releasing local positive supercoiling stress. In this way, REP elements could act as topological insulators that enable (simultaneous-) transcription of genes localized in gene organization structures in which expression of one gene interferes with the expression of its neighbor (e.g. transcription of two convergent genes or differential transcription of a gene downstream of a transcribed gene).

Interestingly, a BIME-2 REP was already shown to be a topological insulator that is able to block transcription induced supercoil diffusion (Moulin et al., 2005). Within this study, Moulin and colleagues created a topological cassette in *Escherichia coli* constituted of a supercoiling inducer (P_R promoter fused to *uidA*) and a reporter (a supercoiling sensitive promotor fused to *lacZ*) to detect local modifications of DNA supercoiling in regions located downstream of transcription units. This enabled them to identify sequences that block transcription induced supercoil diffusion in neighboring regions. They tested various sequence elements that possible influence supercoil diffusion and included different repetitive sequences; the BIME-1 element located at the 3' end of *gyrB*, the *nrdAB* BIME-2 element and two different ERIC/IRU repeats (Enterobacterial Repetitive Intergenic Consensus/Intergenic Repeat Units; (Wilson and Sharp, 2006)). When inserted between the inducer and the topological reporter, BIME-2 *nrdAB* prevented propagation of TI positive supercoils, whereas the BIME-1 element did not prevent modification of local supercoiling in the downstream region (Moulin et al., 2005). The particular BIME-2 used is a DNA gyrase target and the inhibiting effect was attributed

to the cleavage of a DNA fragment by DNA gyrase. These findings confirm our hypothesis that REP elements are actually blocking transcription induced positive supercoiling in the case of BIME-2 but not in the case of BIME-1. It could well be that the net local (negative-) supercoiling was inefficient for the BIME-1 to be in cruciform state and therefore could not act as a buffer reducing the supercoiling stress induced by transcription. This argument implies DNA gyrase, the topoisomerases and particular DNA-binding proteins that affect the local supercoiling state, besides REPs, in the process of relaxing transcriptional interference. Indeed all of these proteins have been related in literature to the role of REPs, as described in the above.

Using the gene expression data from a set of 465 publicly available microarrays (M3D), our study also enabled us to associate REP-related gene-pairs to specific physiological conditions. The strongest indication for a common effect of REPs was the fact that the average difference in expression of the REP associated convergent gene-pairs was lower in all of the experiments than the average difference in expression of the non-REP associated convergent gene-pairs (Fig. 4). The observation is especially striking considering the fact that both the CAI and gene expression of REP associated gene-pairs were significantly higher than their counterparts without REPs (Fig. 3). The array data showed that the effect should not be limited to particular conditions but will carry importance for the *E. coli* cell in any environment it finds itself (Fig. 5). At the same time, every condition is associated to a particular set, or particular sets, of REP associated gene-pairs. However, there seems to be no overarching functional connection between the REP associated gene-pairs other than the connection to a high mutual expression of the pairs.

In conclusion, we hypothesize that REPs can be found in cruciform formation-state on the chromosome and upon transcription of the gene(s) upstream, unfold and thus relieve the effects of transcriptional torsion caused by the expression of upstream gene (mechanism depicted in Fig. 6). We argue that the cruciform state of REPs will differ under different physiological conditions and, in addition, will not be uniform for all REPs on the chromosome. The probability for a REP to be found in cruciform state is related to the (overall-) supercoiling state of the chromosome, which can vary under different conditions (Dayn et al., 1991). For instance, a clear difference in supercoiling state was found between exponentially phase and stationery phase (Balke and Gralla, 1987). Based on our data, we cannot exclude that the observed transcriptional effects are solely due to the binding of proteins such as DNA gyrase (BIME-2 elements) and IHF (BIME-1 elements) to the REP elements and not because of cruciform formation. However, as the overall supercoiling state is affected by these proteins they should affect the action of REPs directly.

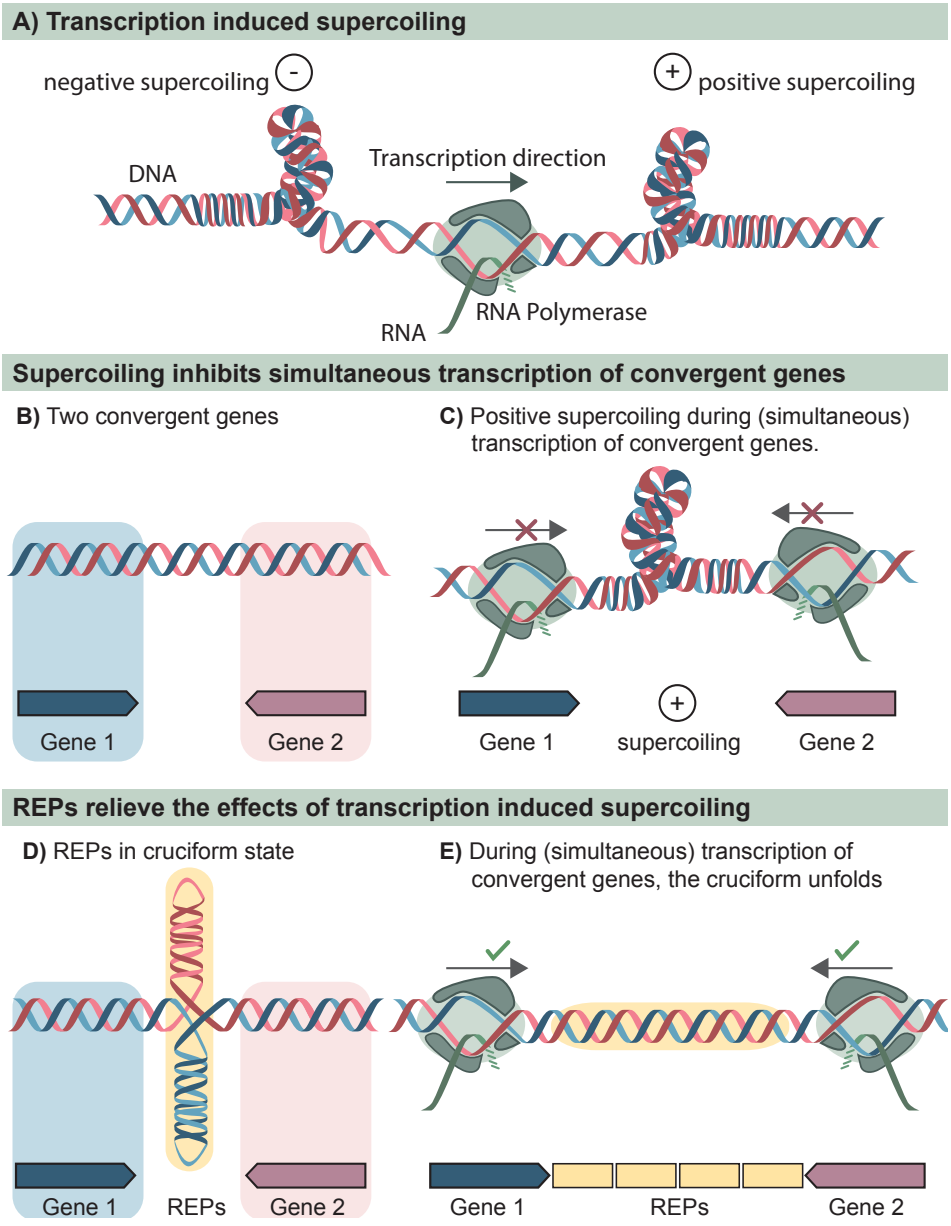


Figure 6. Schematic representation of the REP-enabled relieve of transcription induced positive supercoiling. A) The transcription of a gene induces positive supercoiling in front of the RNA polymerase and leaves negative supercoiling behind**. B) and C)** Simultaneous transcription of two convergent genes without REP elements is inhibited by the transcriptional torsion caused by the two polymerases. **D)** REP elements separating a convergent gene-pair are in cruciform state before the RNA polymerases get closer. **E)** When the two REP associated convergent genes are being transcribed unfolding of the cruciform relaxes the positive supercoiling stress and enables both genes to be transcribed.

Further experimental study could elucidate this important and global but unexplored role of REPs in transcription in *E. coli.* Comparative genome analysis on different *E. coli* strains and/or different REP-containing species could be used to identify those gene-pairs that are consistently REP-associated and those that have a weaker REP association. Analysis of the function of these genes and associations of these strains or species with specific conditions or environments might help to further unravel a functional repertoire of the REP enabled transcriptional network. Moreover, experimental studies using for example supercoiling sensitive promoters could provide a better insight in to the effects of REPs under various conditions and thus under different supercoiling levels. Novel techniques to measure supercoiling such as the use of magnetic tweezers combined with fluorescence (King et al., 2016) might help shed a light on the precise mechanism by which REPs relieve the effects of transcriptional supercoiling stress.

Acknowledgements

This work was part of the BioRange programme of the Netherlands Bioinformatics Centre (NBIC), which is supported by the Netherlands Genomics Initiative (NGI). We thank Martijn van der Pol and Vincent van Deutekom for their contribution to the project.

Supporting information

Supplementary data is freely available online:
https://figshare.com/s/733d34dea145f8304003

References

Aranda-Olmedo, I., Tobes, R., Manzanera, M., Ramos, J.L., and Marques, S. (2002). Species-specific repetitive extragenic palindromic (REP) sequences in *Pseudomonas putida*. *Nucl Acids Res* **30**, 1826–1833.

Bachellier, S., Saurin, W., Perrin, D., Hofnung, M., and Gilson, E. (1994). Structural and functional diversity among bacterial interspersed mosaic elements (BIMEs). *Mol. Microbiol.* **12**, 61–70.

Bachellier, S., Clément, J.M., Hofnung, M., and Gilson, E. (1997). Bacterial interspersed mosaic elements (BIMEs) are a major source of sequence polymorphism in *Escherichia coli* intergenic regions including specific associations with a new insertion sequence. *Genetics* **145,** 551–562.

Bachellier, S., Clément, J.-M., and Hofnung, M. (1999). Short palindromic repetitive DNA elements in enterobacteria: a survey. *Res. Microbiol*. **150**, 627–639.

Balke, V.L., and Gralla, J.D. (1987). Changes in the linking number of supercoiled DNA accompany growth transitions in *Escherichia coli. J. Bacteriol.* **169**, 4499–4506.

Boccard, F., and Prentki, P. (1993). Specific interaction of IHF with RIBs, a class of bacterial repetitive DNA elements located at the 3' end of transcription units. *EMBO J.* **12**, 5019–5027.

Danese, P.N., Pratt, L.A., and Kolter, R. (2000). Exopolysaccharide production is required for development of *Escherichia coli* K-12 biofilm architecture. *J. Bacteriol.* **182**, 3593–3596.

Dayn, A., Malkhosyan, S., Duzhy, D., Lyamichev, V., Panchenko, Y., and Mirkin, S. (1991). Formation of (dA-dT)n cruciforms in *Escherichia coli* cells under different environmental conditions. *J. Bacteriol*. **173**, 2658–2664.

Dayn, A., Malkhosyan, S., and Mirkin, S.M. (1992). Transcriptionally driven cruciform formation in vivo. *Nucleic Acids Res.* **20**, 5991–5997.

Dröge, P. (1993). Transcription-driven site-specific DNA recombination in vitro. *Proc. Natl. Acad. Sci. U. S. A.* **90**, 2759–2763.

Du, X., Wojtowicz, D., Bowers, A.A., Levens, D., Benham, C.J., and Przytycka, T.M. (2013). The genome-wide distribution of non-B DNA motifs is shaped by operon structure and suggests the transcriptional importance of non-B DNA structures in *Escherichia coli. Nucleic Acids Res.* **41**, 5965–5977.

Espéli, O., and Boccard, F. (1997). In vivo cleavage of *Escherichia coli* BIME-2 repeats by DNA gyrase: genetic characterization of the target and identification of the cut site. *Mol. Microbiol.* **26**, 767–777.

Espéli, O., Moulin, L., and Boccard, F. (2001). Transcription attenuation associated with bacterial repetitive extragenic BIME elements*. J. Mol. Biol*. **314**, 375–386.

Faith, J.J., Driscoll, M.E., Fusaro, V.A., Cosgrove, E.J., Hayete, B., Juhn, F.S., Schneider, S.J., and Gardner, T.S. (2008). Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata. *Nucleic Acids Res.* **36**, D866–D870.

Francke, C., Groot Kormelink, T., Hagemeijer, Y., Overmars, L., Sluijter, V., Moezelaar, R., and Siezen, R.J. (2011). Comparative analyses imply that the enigmatic sigma factor 54 is a central controller of the bacterial exterior. *BMC Genomics* **12**, 385.

Gilson, E., Clément, J.M., Brutlag, D., and Hofnung, M. (1984). A family of dispersed repetitive extragenic palindromic DNA sequences in *E. coli. EMBO J.* **3**, 1417–1421.

Gilson, E., Perrin, D., and Hofnung, M. (1990). DNA polymerase I and a protein complex bind specifically to *E. coli* palindromic unit highly repetitive DNA: implications for bacterial chromosome organization. *Nucleic Acids Res.* **18**, 3941–3952.

Gilson, E., Saurin, W., Perrin, D., Bachellier, S., and Hofnung, M. (1991). Palindromic units are part of a new bacterial interspersed mosaic element (BIME). *Nucleic Acids Res.* **19**, 1375–1383.

Gutiérrez-Estrada, A., Ramírez-Santos, J., and Gómez-Eichelmann, M.D.C. (2014). Role of chaperones and ATP synthase in DNA gyrase reactivation in *Escherichia coli* stationary-phase cells after nutrient addition. *SpringerPlus* **3**, 656.

Higgins, C.F., Ames, G.F.-L., Barnes, W.M., Clement, J.M., and Hofnung, M. (1982). A novel intercistronic regulatory element of prokaryotic operons. *Nature* **298**, 760–762.

Khemici, V., and Carpousis, A.J. (2004). The RNA degradosome and poly(A) polymerase of E*scherichia coli* are required in vivo for the degradation of small mRNA decay intermediates containing REP-stabilizers. *Mol. Microbiol.* 770–790.

King, G.A., Peterman, E.J.G., and Wuite, G.J.L. (2016). Unravelling the structural plasticity of stretched DNA under torsional constraint. *Nat. Commun*. **7**, 11810.

Krasilnikov, A.S., Podtelezhnikov, A., Vologodskii, A., and Mirkin, S.M. (1999). Large-scale effects of transcriptional DNA supercoiling in vivo. *J. Mol. Biol*. **292**, 1149–1160.

Liang, W., Rudd, K.E., and Deutscher, M.P. (2015). A Role for REP Sequences in Regulating Translation. *Mol. Cell 58*, 431–439.

Liu, L.F., and Wang, J.C. (1987). Supercoiling of the DNA template during transcription. *Proc. Natl. Acad. Sci.* **84**, 7024–7027.

Ma, J., Bai, L., and Wang, M.D. (2013a). Transcription Under Torsion. *Science* **340**, 1580–1583.

Ma, Q., Yin, Y., Schell, M.A., Zhang, H., Li, G., and Xu, Y. (2013b). Computational analyses of transcriptomic data reveal the dynamic organization of the *Escherichia coli* chromosome under different conditions. *Nucleic Acids Res.* **gkt261**.

Menzel, R., and Gellert, M. (1983). Regulation of the genes for *E. coli* DNA gyrase: homeostatic control of DNA supercoiling. *Cell* **34**, 105–113.

Meyer, S., and Beslon, G. (2014). Torsion-Mediated Interaction between Adjacent Genes. *PLOS Comput Biol* **10**, e1003785.

Miller†, W.G., and Simons, R.W. (1993). Chromosomal supercoiling in *Escherichia coli*. *Mol. Microbiol.* **10**, 675–684.

Moulin, L., Rahmouni, A.R., and Boccard, F. (2005). Topological insulators inhibit diffusion of transcription-induced positive supercoils in the chromosome of *Escherichia coli*. *Mol. Microbiol.* **55**, 601–610.

Newbury, S.F., Smith, N.H., Robinson, E.C., Hiles, I.D., and Higgins, C.F. (1987). Stabilization of translationally active mRNA by prokaryotic REP sequences. *Cell* **48**, 297–310.

Nunvar, J., Huckova, T., and Licha, I. (2010). Identification and characterization of repetitive extragenic palindromes (REP)-associated tyrosine transposases: implications for REP evolution and dynamics in bacterial genomes. *BMC Genomics* **11**, 44.

Oppenheim, A.B., Rudd, K.E., Mendelson, I., and Teff, D. (1993). Integration host factor binds to a unique class of complex repetitive extragenic DNA sequences in *Escherichia coli. Mol. Microbiol.* **10**, 113–122.

Ostrander, E.A., Benedetti, P., and Wang, J.C. (1990). Template supercoiling by a chimera of yeast GAL4 protein and phage T7 RNA polymerase. *Science* **249**, 1261–1265.

Overmars, L., van Hijum, S.A.F.T., Siezen, R.J., and Francke, C. (2015). CiVi: circular genome visualization with unique features to analyze sequence elements. *Bioinforma. Oxf. Engl.* **31**, 2867–2869.

Pontiggia, A., Negri, A., Beltrame, M., and Bianchi, M.E. (1993). Protein HU binds specifically to kinked DNA. *Mol. Microbiol.* **7**, 343–350.

Puigbò, P., Romeu, A., and Garcia-Vallvé, S. (2008). HEG-DB: a database of predicted highly expressed genes in prokaryotic complete genomes under translational selection. *Nucleic Acids Res.* **36**, D524–D527.

Qian, Z., Macvanin, M., Dimitriadis, E.K., He, X., Zhurkin, V., and Adhya, S. (2015). A New Noncoding RNA Arranges Bacterial Chromosome Organization. *mBio* **6**, e00998–15.

Rui, S., and Tse-Dinh, Y.-C. (2003). Topoisomerase function during bacterial responses to environmental challenge. *Front. Biosci. J. Virtual Libr.* **8**, d256–d263.

Sharp, P.M., and Li, W.-H. (1987). The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15**, 1281–1295.

Snoep, J.L., van der Weijden, C.C., Andersen, H.W., Westerhoff, H.V., and Jensen, P.R. (2002). DNA supercoiling in *Escherichia coli* is under tight and subtle homeostatic control, involving gene-expression and metabolic regulation of both topoisomerase I and DNA gyrase. *Eur. J. Biochem. FEBS* **269**, 1662–1669.

Sobetzko, P. (2016). Transcription-coupled DNA supercoiling dictates the chromosomal arrangement of bacterial genes. *Nucleic Acids Res.* **44**, 1514–1524.

Stevenson, G., Andrianopoulos, K., Hobbs, M., and Reeves, P.R. (1996). Organization of the *Escherichia coli* K-12 gene cluster responsible for production of the extracellular polysaccharide colanic acid. *J. Bacteriol.* **178**, 4885–4893.

Teves, S.S., and Henikoff, S. (2014). Transcription-generated torsional stress destabilizes nucleosomes. *Nat. Struct. Mol. Biol.* **21**, 88–94.

Tobes, R., and Pareja, E. (2006). Bacterial repetitive extragenic palindromic sequences are DNA targets for Insertion Sequence elements. *BMC Genomics* **7**, 62.

Tobes, R., and Ramos, J.-L. (2005). REP code: defining bacterial identity in extragenic space. *Environ. Microbiol.* **7**, 225–228.

Tsao, Y.P., Wu, H.Y., and Liu, L.F. (1989). Transcription-driven supercoiling of DNA: direct biochemical evidence from in vitro studies. *Cell* **56**, 111–118.

Wang, J.C. (1985). DNA Topoisomerases. *Annu. Rev. Biochem*. **54**, 665–697.

Wilson, L.A., and Sharp, P.M. (2006). Enterobacterial repetitive intergenic consensus (ERIC) sequences in *Escherichia coli*: Evolution and implications for ERIC-PCR. *Mol. Biol. Evol.* **23**, 1156–1168.

Yang, Y., and Ames, G.F. (1988). DNA gyrase binds to the family of prokaryotic repetitive extragenic palindromic sequences. *Proc. Natl. Acad. Sci. U. S. A.* **85**, 8850–8854.

Zhou, J., and Rudd, K.E. (2012). EcoGene 3.0. *Nucleic Acids Res.* **gks1235**.

Chapter 5

A novel quality measure and correction procedure for the annotation of microbial translation initiation sites

Lex Overmars, Roland J. Siezen, Christof Francke

PloS one, 2015, 10: e0133691.

Abstract

The identification of translation initiation sites (TISs) constitutes an important aspect of sequence-based genome analysis. An erroneous TIS annotation can impair the identification of regulatory elements and N-terminal signal peptides, and also may flaw the determination of descent, for any particular gene. We have formulated a reference-free method to score the TIS annotation quality. The method is based on a comparison of the observed and expected distribution of all TISs in a particular genome given prior gene-calling. We have assessed the TIS annotations for all available NCBI RefSeq microbial genomes and found that approximately 87% is of appropriate quality, whereas 13% needs substantial improvement. We have analyzed a number of factors that could affect TIS annotation quality such as GC-content, taxonomy, the fraction of genes with a Shine-Dalgarno sequence and the year of publication. The analysis showed that only the first factor has a clear effect. We have then formulated a straightforward Principle Component Analysis-based TIS identification strategy to self-organize and score potential TISs. The strategy is independent of reference data and *a priori* calculations. A representative set of 277 genomes was subjected to the analysis and we found a clear increase in TIS annotation quality for the genomes with a low quality score. The PCA-based annotation was also compared with annotation with the current tool of reference, Prodigal. The comparison for the model genome of *Escherichia coli* K12 showed that both methods supplement each other and that prediction agreement can be used as an indicator of a correct TIS annotation. Importantly, the data suggest that the addition of a PCA-based strategy to a Prodigal prediction can be used to 'flag' TIS annotations for re-evaluation and in addition can be used to evaluate a given annotation in case a Prodigal annotation is lacking.

Introduction

The *ab initio* identification of coding sequences is the first step in the annotation of a genome. Various computational methods have been developed to identify coding sequences from Open Reading Frames (ORFs) with low error rate. Automated identification of the Translation Initiation Sites (TISs) associated with the protein-encoding genes has proven to be more difficult. The difficulty probably relates to the fact that the sequence signatures that are associated with the initiation of translation can be diverse. In prokaryotes, the translation of the majority of protein-encoding genes is initiated by the interaction between a short sequence in the 5' untranslated region (5'-UTR) of the mRNA, referred to as the Shine-Dalgarno (SD) sequence (Shine and Dalgarno, 1974), and the 3'-end of the 16S ribosomal RNA. It was observed that the presence of the SD sequence is correlated with a higher expression level (Ma et al., 2002). Similarly, the presence of the SD sequence correlated with the occurrence of an AUG codon as the translation start (Ma et al., 2002). Nevertheless, the SD sequence is not absolutely required as it was found that many, and even some highly translated, mRNAs lack a (recognizable) SD sequence (Skorski et al., 2006). So far, two alternative (i.e., SD-independent) mechanisms of translation initiation have been identified (Nakagawa et al., 2010). The first SD-independent mechanism involves ribosomal protein S1 (RPS1), which interacts with the 5'-UTR to initiate translation (Komarova et al., 2005). The second mechanism involves the 70S ribosome as a whole, which can interact directly with leaderless genes (genes without a 5' UTR) and uses an N-formyl-methionyl-transfer RNA to initiate translation (Moll et al., 2002; Udagawa et al., 2004). The start codon is assumed to be the most important signal for the translation of leaderless genes. Analysis of 162 completed bacterial genomes showed that the number of genes not preceded by an SD-sequence is highly variable between bacteria, where the reported number varies between 9.2% and 88.4% (Chang et al., 2006; Zheng et al., 2011).

Currently the most widely used gene-calling tools are GLIMMER3 (Delcher et al., 2007) and Prodigal (Hyatt et al., 2010). Other tools include MED2.0 (Zhu et al., 2007), GeneMarkHmm (Lukashin and Borodovsky, 1998) and EasyGene (Larsen and Krogh, 2003). The former tools predict coding sequences with relative low error rates for genomes of well-studied organisms. Nevertheless, the annotation of genes in high-GC-content genomes using these tools is more challenging, since the genomes contain fewer random stop codons leading to longer Open Reading Frames (ORFs) and more mistakes (Hyatt et al., 2010). Three main approaches are in use to improve upon a given TIS annotation. These are essentially based on: i) post-processing of initial predictions; ii)

comparative genomics; and iii) combining multiple predictions. The related tools commonly start from existing genome annotations or genes identified by the before-mentioned prediction tools. For instance, TICO (Tech et al., 2005) was developed to improve the accuracy of TIS annotation by performing an unsupervised classification of strong-TIS and weak-TIS sequences. Similarly, various resources such as ProTISA (Hu et al., 2008a) and SupTISA (Hu et al., 2008b) have accumulated (post-processed) predictions from different sources. In ORFcor, orthologous sequences are used to identify and correct inconsistencies in the gene and TIS annotation (Klassen and Currie, 2013). Likewise, Genome Majority Voting was used to assign TISs based on groups of orthologous sequences (Wall et al., 2011). The pipeline GenePRIMP (Pati et al., 2010) was developed to improve the gene prediction of bacterial genomes and to report anomalies including inconsistent start sites, and missed and split genes. Multiple gene-prediction methods have been combined to improve the accuracy of gene and TIS annotation (Ederveen et al., 2013; Shah et al., 2003; Yada et al., 2003; Yok and Rosen, 2011). It was found that the application of a specific path in the combination of predictors can provide a gain in sensitivity while maintaining a high specificity in gene prediction (Ederveen et al., 2013). Nevertheless, a recent comparison of the various available prediction tools and pipelines indicated that the best performers achieved a maximal TIS prediction accuracy of around 90% for a typical genome (Hyatt et al., 2010). Moreover, the addition or combination of tools did not often lead to an improvement in the estimated quality above 90%.

Different types of errors are commonly introduced by computational gene calling and annotation methods. First, true coding regions can be overlooked. However, the percentage of missed genes is estimated not to exceed 5-10% (Poptsova and Gogarten, 2010). Second, some predicted genes do not represent a true coding sequence (Ederveen et al., 2013; Hyatt et al., 2010). Third, the assignment of the correct start codon (i.e., the translation initiation site (TIS)) can be erroneous. Bakke and colleagues (Bakke et al., 2009) evaluated the performance of three automated genome annotation services for the annotation of the archaeon *Halorhabdus utahensis*, namely: IMG (Markowitz et al., 2006), RAST (Aziz et al., 2008) and the J. Craig Venter Institute (JCVI) Annotation Service. There appeared to be considerably more agreement concerning the identified translation stop codons (90% shared) than concerning the annotated TISs (48% shared) between the three services. The inconsistency in TIS annotation was also highlighted by another study, in which it was shown that 53% of the orthologs among 5 *Burkholderia* genomes have inconsistently annotated TISs in RefSeq (Dunbar et al., 2011). The incorrect annotation of TISs can flaw different types of genome analysis such as: the (automated) identification of regulatory sequences, the construction

of reliable phylogenetic trees for homologous genes/proteins, the function annotation of the gene product and the prediction of the subcellular location of the gene product.

An important limitation in *de novo* gene prediction is the need for reference data-sets with correctly identified TISs to test the quality of annotations. Unfortunately, large sets of translated proteins where the N-terminus has been experimentally verified are scarce (Smollett et al., 2009). A frequently used dataset of verified protein sequences is available for *Escherichia coli* K12 MG1655 from EcoGene (Zhou and Rudd, 2013). The translation start sites (926) in this dataset are reported to be experimentally determined using N-terminal protein sequencing. In this paper we present a strategy that avoids the need of reference datasets to assess the accuracy of genome-wide TIS annotation. The strategy involves a comparison between the distribution of alternative TISs around the annotated TISs within a genome, and an expected distribution that can be calculated based on simple and transparent criteria. Such a comparison appeared to provide an intrinsic quality metric for genome-wide TIS-prediction accuracy. We have evaluated the TIS quality for all sequenced genomes and found that the majority was reasonably well annotated, but a substantial minority (~13%) clearly needs to be improved.

In addition, we have developed an iterative Principle Component Analysis (PCA)-based strategy that uses the sequences surrounding all putative TIS for a gene, to identify the most likely TIS. The strategy neither involves training nor reference data, and is not based on any additional assumptions. It can thus be used for any genome. We have implemented the strategy and assigned TISs to all genes for a set of 277 representative bacterial genomes. Comparison of the TIS annotation for the *E. coli* K12 MG1655 genome as obtained with the PCA-based method to the annotation obtained using the standard tool Prodigal revealed a clear advantage of using both methods simultaneously.

Results

An inherent metric to assert the quality of genome-wide gene-predictions

We identified all alternative in-frame translation initiation sites (TISs) for the annotated genes in the complete archaeal and bacterial genomes available via NCBI in January 2013 (see methods). We plotted the distribution of the position of all alternative TISs with respect to the annotated TISs (dataset available at Figshare; http://dx.doi.org/10.6084/m9.figshare.1460717). For the well-studied bacterium *E. coli* K12 MG1655 we found a characteristic distribution, where the number of alternative TISs in the coding part of the gene was reasonably constant and where the number decreased nearly exponentially upstream of the annotated start (Fig. 1A). Furthermore we observed that in *E. coli* K12 MG1655 the first 5 to 10 codons of the coding sequence showed a relative underrepresentation of alternative TISs. In fact, the genomes of other well-studied bacteria such as *Bacillus subtilis* str. 168, *Lactobacillus plantarum* WCFS1, *Listeria monocytogenes* EGD-e, *Pseudomonas putida* KT2440, *Mycobacterium tuberculosis* M37Rv and *Salmonella typhimurium* LT2, showed a very similar distribution (Fig. S1). Moreover, the distribution of *Bacillus subtilis* str. 168 showed a characteristic peak of alternative starts 3 codons upstream from the annotated TIS. The same peak was observed for the other species of the phylum *Firmicutes* (Fig. S1 panels D,C,F). We have also determined the distribution of alternative TISs in *Saccharomyces cerevisiae* S288C (Fig. S1, panel H). The distribution of alternative TISs in this eukaryote appeared highly similar to the ones of the well-studied bacteria. At the same time, we found that a considerable number of genomes (~13%) showed a dissimilar distribution. The dissimilar distributions were of two types: i) a distribution that suggested alternative TISs upstream of the annotated TISs were absent (Fig. 1B); and ii) a distribution that suggested that there was a relatively low probability to find a stop codon upstream of a TIS (Fig. 1C). The former distributions commonly showed a peak of alternative TISs in the first 10 codons of presumed coding sequence. The latter distributions showed a peak of alternative TISs around 30 to 120 nucleotides upstream of the annotated TISs.

While studying the relative positions and frequencies of alternative in-frame TISs we realized that the overall distribution of the alternative start codons should in fact be a very good measure of TIS annotation quality. The distribution should ideally follow the distributions as found in the well-studied organisms, which all displayed a near exponential decrease upstream of the annotated TISs. We calculated the expected distribution for every individual genome given two simple premises: i) the probability of finding alternative

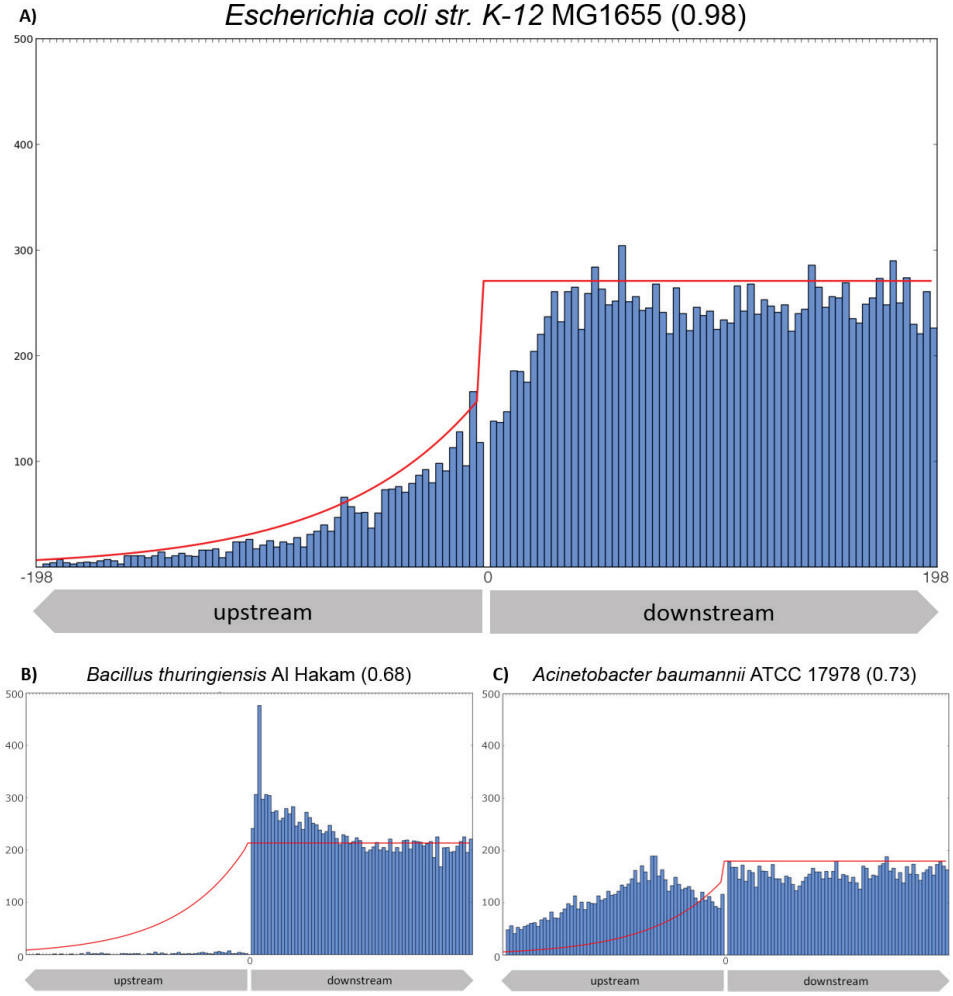


Figure 1. The three typical distributions of alternative start codons found for genomes in the NCBI RefSeq database. A) The distribution of alternative starts in *Escherichia coli* K12 MG1655; **B)** *Bacillus thuringiensis* str. Al Hakam; and **C)** *Acinetobacter baumannii* ATCC 17978. For all ORFs that included an annotated gene and TIS, the total number of alternative start codons for each codon position relative to the annotated translation start were counted. The green line represents the expected distribution as determined using formula 1 In genomes that adhere to figure 1A the observed and expected distribution are alike, whereas ifor genomes that adhere to B or C the observed distribution of alternative start codons given the annotation is clearly deviating from the expected distribution (green line)A comparison of the observed and expected distribution provides an inherent quality measure for genome-wide gene-prediction accuracy.

TISs in coding sequence is constant on average and likewise in non-coding sequence; and ii) the probability of finding an in-frame stop-codon upstream of the TIS is constant on average. To take variation in AT or GC content in account we calculated genome specific in-frame alternative start codon

frequencies and genome specific stop codon frequencies (see methods). The difference between the observed and the calculated distribution could then be used directly as a quality measure of TIS annotation. To probe the difference between the given and the expected distribution we decided to use a correlation measure since such a measure should be relatively insensitive to deviations at particular codon positions. We calculated a Spearman correlation coefficient between the given and expected distribution for all sequenced genomes in the NCBI RefSeq database of January 2013. The calculated correlations of both the upstream and the complete distribution for all analyzed genomes can be found in Table S1. We observed striking differences between the given and expected distribution of alternative TISs for various genomes and found that these differences were more prominent in the upstream region. We therefore decided to use only the upstream

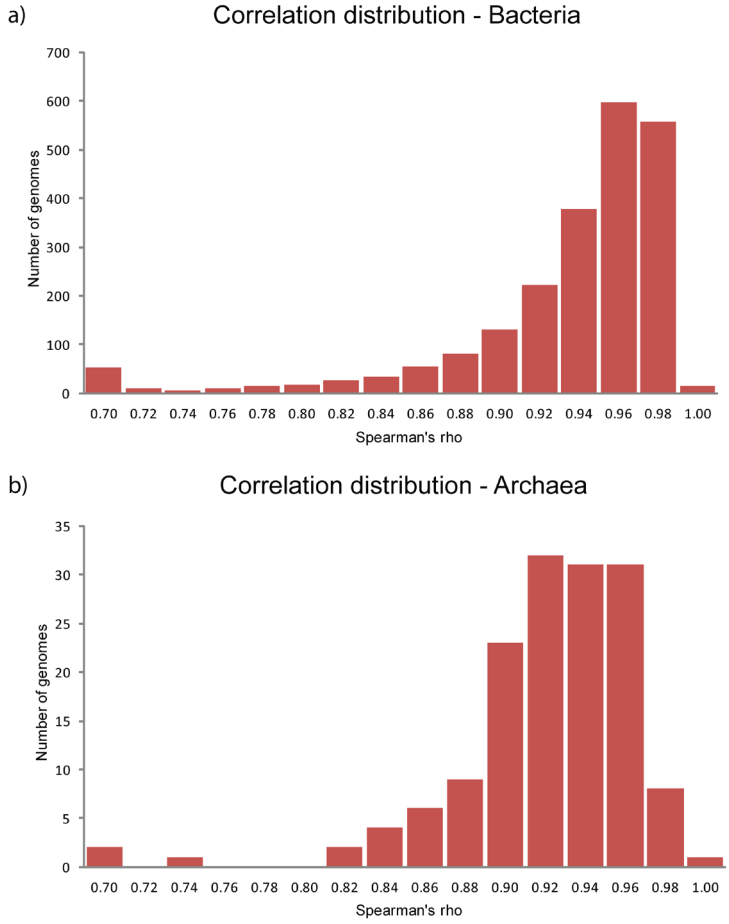


Figure 2. Correlation coefficients between observed alternative start frequencies and expected alternative start frequencies for microbial genomes. A) Spearman's rho coefficients for all bacterial RefSeq genomes with > 500 ORFs. **B)** Spearman's rho coefficients for all Archaeal RefSeq genomes with > 500 ORFs.

correlation for comparison. Based on our two simple premises the calculated correlation coefficient should be a good measure of TIS annotation quality. Indeed the genomes known to be well annotated, like those of *Escherichia coli* K12 MG1655 and *Bacillus subtilis* str. 168, showed a high correlation coefficient (0.98 and 0.98, respectively). On the other hand, the distributions with a low correlation coefficient coincided with the a-typical distributions of alternative TISs similar to those depicted in Fig. 1B and Fig. 1C. In case we used a correlation coefficient of >0.85 for genomes from 500-1500 ORFs and >0.9 for genomes from 1500 ORFs (see discussion) as indicative, the majority of genomes would be qualified as appropriately annotated (Fig. 2). We found that in 88% of the bacterial genomes and in 71% of the archaeal genomes the TIS annotation quality measure was above the threshold (bacteria: 1936 of 2205, archaea 107 of 150).

Factors that affect the quality of the annotation of TISs

The correlation between the given distribution of alternative TISs and the expected distribution was calculated per genome. An important consequence of this way of calculation was that it abolished the need for a reference gene-set and allowed a direct comparison of TIS annotation quality between genomes of varying GC content. For instance, we used the correlation measure to test the change in TIS annotation quality throughout the years. It has been assumed that the quality of the gene calling procedure, which includes the identification of TISs, has decreased in time due to the relative decrease in the number of manually curated annotations and the strong increase in the number of automated annotations (Richardson and Watson, 2013). Contrary to expectation, a comparison of the alternative TIS distribution correlation coefficients against the year of publication (Fig. 3A) did not show such a trend.

Other factors, including GC-content, have also been proposed to be correlated to TIS annotation quality. We found that GC-content indeed correlated with the alternative TIS distribution correlation coefficients. Both high GC-content (>60%) and low GC-content (<40%) genomes showed a relatively low correlation between the given and expected alternative TIS distribution (Fig. 3B). Using a Fisher exact test we determined that the occurrence of above-average quality gene annotation (correlation score > 0.85/0.9) for low and high GC-content genomes compared to the occurrence of high quality gene annotation in moderate GC-content genomes was significantly lower in all cases (p-value 0.0001 or smaller; table S2). We also observed a decreased correlation score for particular phyla (Fig. 3C). However, the effect of phylum could be explained completely by the difference in the GC-content of the species within the phyla.

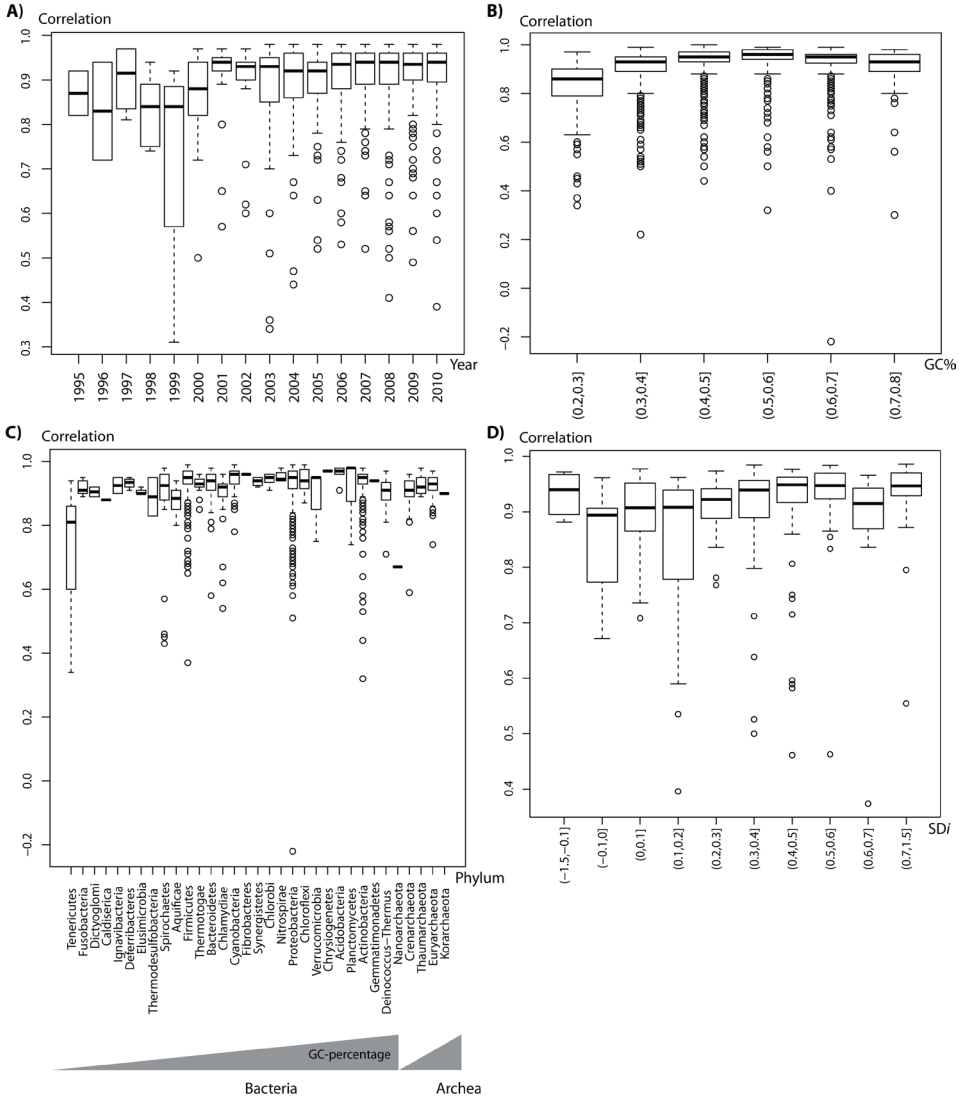


Figure 3. Effects of year of sequencing, GC-content and taxonomy on TIS-prediction accuracy. The boxplots show the distribution of the calculated correlation values (between the observed and expected distribution of alternative TISs) (*Y* axis) for: **A)** all bacterial and archaeal RefSeq genomes grouped by year of sequencing (NCBI Bioproject data; (Krug et al., 2013)); **B)** The RefSeq genomes grouped into 6 bins according to their GC%; **C)** The RefSeq genomes grouped according to phylum; and **D)** 277 selected bacterial and archaeal genomes with varying SD-index (proportion of Shine-Dalgarno sequence-preceded genes) (Nakagawa et al., 2010).

The number of genes preceded by a Shine-Dalgarno sequence within the genomes was another factor that we considered. We found that genomes with a lower SD presence on average had a lower TIS annotation quality (Fig. 3D and Fig. S2). Yet, this tendency was not uniform as the group with the lowest SD presence (mostly *Cyanobacteria* and *Bacteroidetes*) had a relative higher

TIS annotation quality. Vice versa, the group with a SD presence between 0.6 and 0.7 (mostly *Firmicutes* and *Proteobacteria*) had a relative decreased TIS annotation quality. Finally, we observed small differences in TIS annotation quality between genomes sequenced and annotated by the large sequencing centers (Fig. S3).

The use of Principal Component Analysis to identify TISs

Although the observed distributions of alternative TISs overall followed the expected distribution well, this was much less true for the sequence region directly upstream and downstream of the annotated TISs. For example, in *B. subtilis* str. 168 and *E. coli* K12 MG1655 a relative low number of alternative TISs were found in the first codons of the coding sequence, and peaks of alternative TISs were present in the upstream codons preceding the annotated TISs (Fig. 1A and Fig. S1). In fact, the observed deviations should be expected in case recognition of the TIS requires a specific sequence signature. As a consequence, we reasoned, the true TISs should be separable from the alternative TISs based on the signature. Furthermore, the sequence signature related to translation initiation should stand out when the variability of the sequence directly upstream and downstream of potential TISs would be analyzed. The upstream and downstream sequences of all potential TISs for every annotated ORF in a particular genome were therefore converted to binary vectors (as described in the methods). A PCA was initiated using the vectors corresponding to the three longest potential gene-products for every ORF. Given the available data on model organisms (e.g., the *E. coli* reference set in EcoGene 3.0 (Zhou and Rudd, 2013)) the resulting set of vectors should represent a substantial number of true TISs (estimated number >20%) whilst ensuring that the majority of vectors represented false TISs (>66%). To enrich the set with true TIS corresponding vectors we iterated the PCA procedure (see methods). We found that the analysis converged within ten iterations for every genome analyzed. We thus have formulated an iterative PCA-based procedure to separate true TISs from alternative TISs (Fig. S4).

We have applied the PCA-based procedure for *E. coli* K12 MG1655 and iteratively scored all potential TISs. We have included a table containing scores for the 5 best scoring TISs per gene for *E. coli* K12 MG1655 (table S3). The scripts that were used have been made available via Github (see supplementary information) and can be used to evaluate the TIS annotation of any genome. When we employed the simplest assignment scheme, that is using the highest score achieved on principle component I during the iterations to discriminate the true TIS, ~85% of the TISs in *E. coli* K12 MG1655 were assigned identically when compared to the original annotation

in the NCBI RefSeq database (Table 1). The majority of the non-compliant TIS annotations were located downstream of the annotated TIS, with a clear peak at the first downstream codon (Fig. 4). To evaluate the optimal combination of upstream and downstream sequence lengths we performed the procedure

Table 1. TIS annotation for *E. coli* K12 MG1655. The NCBI RefSeq file contained 4141 annotated genes. The position of the TISs was compared between the PCA-based prediction, the Prodigal-based prediction and the RefSeq annotation. Recently, the EcoGene annotation has been updated and 13 TISs have been adjusted (b0259, b0552, b0656, b1994, b2030, b2192, b3218, b3505, b4543, b2803, b1331, b2982 and b3093). The adaptations were compared to the PCA-based and Prodigal-based predictions.

Annotation consistency#	Total	Verified Set	Ecogene Adjusted	Ecogene adjustment
RefSeq = PCA = Prodigal	83% (3418)	88.4% (811)	1	12 nt upstream (b4543)
(RefSeq = Prodigal) ≠ PCA	9.8% (406)	7.8% (71)	0	
(Refseq = PCA) ≠ Prodigal	4% (173)	2.2% (20)	0	
RefSeq ≠ (PCA = Prodigal)	2% (88)	1.4% (13)	12	All in agreement with PCA=Prodigal
Refseq ≠ PCA ≠ Prodigal	1% (54)	0.2% (2)	0	

The majority of TISs that are different in the PCA-based and Prodigal-based annotation are located close to the RefSeq TIS. For the PCA-based predictions: 548 were not in agreement with RefSeq, 199 of these where within 30 nt distance and 56 at 3 nt distance; For the Prodigal predictions: 241 (6%) were not in agreement with RefSeq (and 74 (2%) were missed): 96 of these were within 30 nt distance and 30 at 3 nt distance.

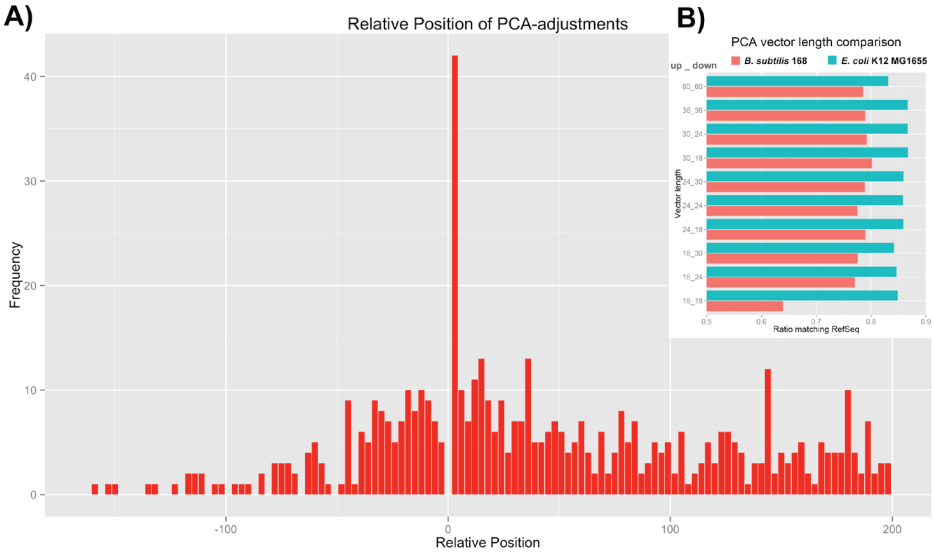


Figure 4. A) The relative position of PCA-based TIS annotations that deviate from the RefSeq annotation for *E. coli* MG1655. B) The effect of sequence vector length on the number of matching PCA-based and RefSeq TIS annotations in *E. coli* K12 MG1655 and *B. subtilis* 168 . The following vector lengths were compared (denoted as: length upstream in nt. and length downstream in nt.): i) 60 & 60, ii) 36 & 36, iii) 30 & 24, iv) 30 & 18, v) 24 & 30, vi) 24 & 24, vii) 24 & 18, viii) 18 & 30 ix) 18 & 24 and x) 18 & 18.

for the phylogenetically distant model organisms *E. coli* MG1655 and *B. subtilis* str.168 using sequence vectors of varying lengths (see methods; the vector lengths are given in the caption to Fig. 4). We found that for the reference genomes the best annotation results were obtained with sequence vectors that represented 30 nt upstream and 18 nt downstream of the annotated TIS (Fig. 4B). We therefore decided to use sequence vectors of this length to annotate *E. coli* MG1655 (Table S4), *B. subtilis* str. 168 and a selected set of bacterial genomes.

We applied the PCA-based annotation strategy to the 277 genomes selected by Nakagawa and colleagues (Nakagawa et al., 2010). The selection of genomes was made to provide a balanced representation of the bacterial and archaeal kingdom in terms of number of genomes per phylum. We found that using the iterative PCA procedure and simple scoring the calculated correlation of the distribution of alternative TISs with respect to the annotated start codon improved significantly for genomes with a poor correlation (and hence a poor TIS annotation quality) (Fig. 5A). Only some of the high quality TIS annotations became slightly worse when applying our simple PCA-based ranking (see quality scores in Table S5). Moreover, the quality of the PCA adjusted TIS annotation appeared to depend hardly on GC-content. An average quality measure of 0.91 was achieved, compared to an average score of 0.90 for the RefSeq annotations. The original genomic distributions of alternative TIS are supplied online (dataset available at Figshare; http://dx.doi.org/10.6084/m9.figshare.1460717).

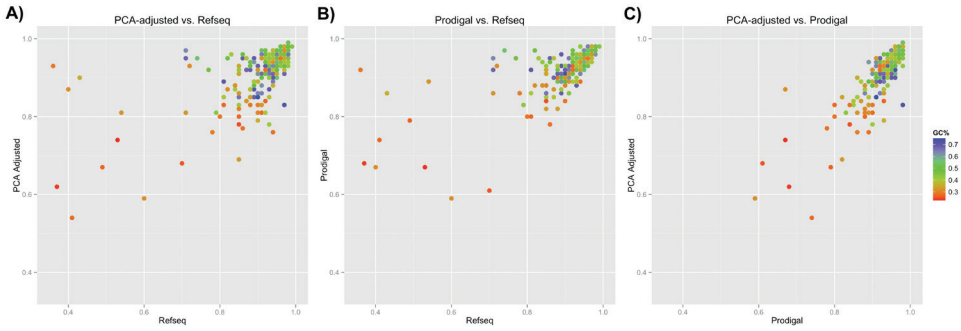


Figure 5. A comparison of TIS prediction accuracy between RefSeq, PCA-based and Prodigal annotation. Scatterplot of the correlation between observed alternative start codon frequencies and expected alternative start codon frequencies (i.e., the TIS annotation quality measure) for both the original TIS annotation as found in the RefSeq database (*Y* axis) and the adjusted annotations (*X* axis) based on **A)** our iterative PCA pipeline and **B)** Prodigal. **C)** Scatterplot for PCA-based annotation versus Prodigal. The color scale represents the GC% of the corresponding genome (blue: high, green: average, red: low)

For comparison, we also performed a *de novo* ORF annotation of the 277 bacterial- and archaeal- genomes using Prodigal (Hyatt et al., 2010). Prodigal achieved a similar increase in the TIS annotation quality score with an average score of 0.92 (Fig. 5B). Our PCA-based method and Prodigal showed a good correspondence in alternative TIS distribution correlation coefficients, where Prodigal performed only slightly better (Fig. 5C and Table S5). Moreover, we found that for almost all genomes Prodigal did not provide an ORF and TIS annotation for all ORFs of the NCBI annotation file. For various genomes the number of ORFs without matching Prodigal annotation exceeded 10% of the total.

To evaluate the differences in performance between the PCA-based TIS annotation and the TIS annotation by Prodigal in more detail, we compared the PCA-based annotation and the Prodigal annotation with the RefSeq annotation for the model organism *E. coli* K12 MG1655 and with the TIS annotation present in the well-curated Ecogene database (see Table 1). For the majority (82.6%) of genes the TIS annotation based on the PCA-based method and on Prodigal corresponded to the RefSeq annotation. For 406 (9.8%) genes the TIS annotation using the PCA-based method conflicted with the Prodigal and Refseq annotation. Vice versa, for 173 genes (4.2%) the PCA-based prediction was consistent with the RefSeq annotation but conflicted with that of Prodigal. These included 74 genes that were not called using Prodigal (e.g. no matching stop codon was found). Interestingly, for 88 genes (2.1%) the TIS annotation of both the PCA–based method and Prodigal were identical but conflicted with the RefSeq annotation. Moreover, we observed that for a large number of the genes in the latter group the distance between the annotated TISs was less than 30 nt (see Table 1 and Fig. 4). Only in 54 cases (1%) all three annotations disagreed.

Discussion

Two recent comparisons of the common gene identification algorithms showed that the algorithms mostly agree on the location of the genes but quite often provide an inconsistent positioning of the TISs (Dunbar et al., 2011; Ederveen et al., 2013). Due to the availability of only a limited number of reliably curated genome annotations, TIS identifications might be biased. In fact, even for model organisms the number of datasets containing experimentally validated TISs is scarce (Smollett et al., 2009). The effect of bias on prediction quality is potentially underestimated given the fact that the identification algorithms have mostly been benchmarked using the same reference data.

We argue that quantification of the similarity between an observed (genome-wide) and the expected distribution of alternative TISs with respect to the annotated TISs provides an inherent measure of TIS annotation quality. The measure solely depends on the genome sequence that is being analyzed and is therefore reference independent. It is easy to establish, compare and interpret. We have implemented the proposed quality measure and found that in all the genomes assumed to have a high quality TIS annotation (i.e., reference genomes used in other papers) the observed distribution of alternative TISs corresponded well with the distribution that we calculated based directly on expected triplet frequencies. Therefore, the correlation between the observed distribution of alternative TISs and the expected distribution of alternative TISs appears indeed to be a good measure for TIS annotation quality. Moreover, the outcome of a similar analysis of TIS distribution for all chromosomes of the reference yeast *Saccharomyces cerevisiae* S288c (correlation: 0.98) suggests the quality measure can as easily be applied to eukaryotic genomes.

Our TIS distribution-based correlation measure was used to score TIS annotation quality for completely sequenced bacterial and archaeal genomes. Although a score of 1.0 reflects a perfect correlation, it must be noted that a somewhat smaller score probably already reflects the 'perfect' score as the occurrence of some abnormal TIS sites could decrease the correlation slightly. Moreover, the number of alternative upstream TISs is in all cases relatively low –the number was lower than 150 directly upstream of the annotated TIS and decreased to zero within ~200 nucleotides for all genomes that were studied- and the related distribution should be relatively noisy as a consequence, thus reducing the correlation. Indeed, the well-annotated prokaryotic genomes of *E. coli* MG1655 and *B. subtilis* str.168 and the eukaryotic genome of *S. cerevisiae* S288c showed such a slightly reduced supposed optimal correlation of 0.98. We have used a correlation score >0.9 as indicative of appropriate TIS annotation. A correlation value >0.9 was obtained for all genomes with more than 1500 ORFs in the set of 277 selected genomes using either the original RefSeq TIS annotation, or the PCA-based annotation, or the Prodigal annotation (Table S5). We observed that for smaller genomes the spread in correlation values became somewhat larger (Fig. S5). For genomes comprising 500 to 1500 ORFs therefore a correlation higher than 0.85 can be used as indicative of appropriate annotation.

We found that the genomes with a relative poor TIS annotation quality (500-1500 ORFs and score ≤0.85; >1500 ORFs and score ≤0.9) comprised about 13% of the genomes deposited in the Refseq database in January 2013. The TIS annotation of archaeal genomes is relatively more frequently of lower

quality. For some genomes a very atypical distribution of alternative TIS was found, resulting in a very low annotation quality in the NCBI genome database. These genomes include for instance *Rhodospirillum photometricum* DSM 122 (0.22), *Rothia mucilaginosa* DY-18 (0.32), *Clostridium tetani* E88 (0.34) and *Borrelia turicatae* 91E135 (0.43). We have checked the corresponding publications to find abnormalities in gene-calling procedures and genome sequence characteristics and we found that indeed less common gene-calling procedures were used. Also genomes that are sometimes used as reference genomes were found to have a questionable annotation quality. For instance, the TIS distribution in *Streptococcus pneumoniae* R6 had a correlation of 0.83, whereas the correlation for other *Streptococcus pneumonia* genomes was >0.9. These observations underline the necessity to be careful in selecting a reference genome.

We have used the correlation measure to evaluate several factors that have been proposed to affect the TIS annotation quality. Contrary to expectation, we observed no correlation between the year the genome was published (annotated) and TIS annotation quality (Fig. 3). This might well be related to the substantial increase in the quality of the annotation tools during the last decade. The sequencing center appears to have a small effect. Surprisingly, also the percentage of genes preceded by an SD-sequence in a genome does not seem to affect the TIS annotation quality much. In contrast, we found that genomes with low and high GC-content showed a significantly decreased TIS annotation quality. Thus low GC and high GC-content appear to be more problematic where the proper annotation of TIS by 'traditional' means is concerned.

Interestingly, the abundance of alternative TIS in the direct context of annotated TISs was aberrant. For example, in most, if not all, *Firmicutes* genomes we observed a characteristic peak of alternative TISs located 9 nucleotides upstream of the annotated TIS. The ribosomal binding site in these genomes explains this characteristic deviation. The full Shine-Dalgarno motif (AGGAGGU) needs only to be followed by a G to attain a GUG alternative start-codon. Genomes that belong to the *Firmicutes* phylum were among those reported to have the most genes preceded by a Shine-Dalgarno motif (up to 92%) (Pallejà et al., 2009). In line with this, we observed that the majority of alternative starts in the observed peaks in the *Firmicutes* genomes are indeed GUG-codons. At the same time, relatively few alternative TISs were observed in the first codons of the coding sequence. Recent analysis of eukaryotic coding sequences also showed low numbers of AUG codons in the first 5-11 codons following the TIS. The low numbers were attributed to the prevention of translation of alternative genes (Zur and Tuller, 2013). Our observations for

bacterial genes suggest that the low abundance of alternative start codons in the first part of the coding sequence is a universal trait of genes. Further work is under way to explore the possible biological relevance of the variability in codon abundance upstream and downstream of the TISs.

We have formulated a PCA based TIS scoring strategy and applied it to distinguish the true TISs from alternative TISs. An important advantage of the strategy is that it is self-organizing and that it thereby circumvents the need for reference data and knowledge of the characteristics of the genome that is analyzed. Thereby, every genome can be analyzed in the same way without impairing the overall quality. The PCA output can be used directly to manually assess the TIS annotation of individual ORFs.

A simple scoring scheme was applied to utilize the scores from the different iterations of PCA in an automated manner. We show that by applying the simple scoring scheme the TIS annotation quality of many of the relatively bad scoring genomes can be improved (Fig. 5A). Moreover, The TIS annotation quality score observed after the PCA-based re-annotation of 277 representative bacterial genomes supports the reduced dependency between genome characteristics and TIS annotation quality, when compared to other predictors. We compared the quality of the PCA-adjusted TIS annotations with the ones derived from Prodigal and found that both methods improve the gene annotations in various genomes.

To evaluate the quality of the individual TIS annotations provided by the PCA-based strategy we compared the TIS annotation in *E. coli* K12 MG1655 provided by NCBI's RefSeq database and by the Ecogene database (Zhou and Rudd, 2013), with the annotations calculated using our PCA-based strategy and Prodigal. We found that PCA-based- and Prodigal annotations were in agreement for the majority of genes (83%). The agreement is much better than the overall agreement in TIS annotation observed by (Bakke et al., 2009) between different much-used annotation services, and the overlap observed for *Burkholderia* orthologs (Dunbar et al., 2011). Importantly, the algorithms and information employed in our PCA-based strategy may be viewed as predominantly independent from the algorithms and information used by Prodigal. Whereas, Prodigal and most annotation procedures rely on the use of Hidden Markov Models and Dynamic Programming algorithms to score the TISs (Wang et al., 2013) and use reference data and the whole genome sequence to make sequence models, our procedure is devoid of such models and solely employs the sequence surrounding potential TISs and self-organizes those sequences using PCA. Therefore, the PCA-based strategy adds valuable independently obtained information to the Prodigal annotation.

In *E. coli* K12 MG1655, we found that in 2% of the cases the PCA-based and Prodigal predictions complied but were different from the Refseq annotation. Although the Refseq TIS annotation of *E. coli* is used as a reference, it does contain mistakes. For instance, in a recent update of the Ecogene database (Ecogene 3.0 (Zhou and Rudd, 2013)) 13 adjustments (with respect to the Refseq annotation) were made in the annotation. In fact, all of these were made to genes for which the PCA-based and Prodigal TIS predictions agreed, further suggesting that a compliance between the PCA-based and Prodigal annotation is a strong indicator for a correct TIS annotation. The fact that for the experimentally verified set of genes the correct prediction rate is even higher in the case of compliancy further supports the assertion. Moreover, recently the *fes* gene (b0585) was removed from the experimentally verified set (Zhou and Rudd, 2013) because the N-terminus of the encoded protein that was reported in literature before (Pettis et al., 1988) was found 26 amino acids too short in a shotgun MS experiment (Krug et al., 2013). The related TIS annotation, which was located 78 nt upstream of the TIS annotated in the Refseq database, was assigned correctly by both our PCA-based method and Prodigal. The incorrect annotation was corrected in the latest update of the *E. coli* K12 MG1655 Refseq record.

The above implies that consistency between a PCA-based annotation and Prodigal is a good indicator of proper TIS annotation and suggests that it will be useful to manually evaluate the TIS annotation in case the Prodigal/PCA-based annotation disagrees with the existing annotation. A difference between the PCA-based and Prodigal prediction would be another reason to 'flag' the annotation for manual curation. We found that for 14% of the *E. coli* genes either the PCA-based (406; 10%) or the Prodigal (173; 4%) TIS annotation were not in agreement with the RefSeq TIS annotation. Considering the numbers, adopting the Prodigal annotation instead of the PCA-based annotation would lead to a better annotation in the case of *E. coli*. However, the performance of both prediction strategies depends clearly on the genome that is being annotated as is implied by the data in Fig. 5. Therefore a preference for one method over the other cannot be generalized. In addition, we observed that for a large number of the genes with a different TIS annotation between the PCA-based method and Prodigal the distance between the annotated TISs was less than 30 nt and was actually peaking at 3 nt (a single codon difference). Given the self-organizing nature of the PCA-based method it could be that these TIS in fact have been incorrectly annotated in RefSeq. Importantly, the Refseq and Ecogene database also contained a number of annotated genes that where not found using Prodigal. For 74% of these genes (55 out of 74) the PCA-based TIS annotation was the same as the one found in the reference database. This implies that the PCA-based TIS

scoring strategy can be used to evaluate the TIS annotation for ORFs that are not recognized by Prodigal or other conventional gene prediction methods. In fact, we found that for various genomes Prodigal missed more than 10% of the ORFs reported in the RefSeq database (Table S5).

The percentage of genes for which the PCA-based prediction was different from the Prodigal prediction and both different from the RefSeq annotation was around 1%. This small number implies that a combination of the sequence information used in the PCA-based method and the algorithms used in Prodigal together must capture most of the properties of the transcript sequence that determine the location of translation initiation sites in *E. coli.* In fact, the deviant TISs might be interesting to investigate in more detail because of their atypical character and thus because of a potentially alternative translation initiation mechanism.

Conclusion

The newly defined distribution-based score for TIS annotation provides a powerful tool for the assessment of TIS annotation quality because it can be employed on any genome sequence without the need for a reference. We have evaluated the TIS annotation quality of the complete bacterial genomes present in the NCBI RefSeq database and found that a significant portion of genomes (~13%) has a questionable TIS annotation. To improve the quality of the genome annotation data in the public domain we therefore would consider it valuable that the TIS annotation quality is assessed before researchers publish their genome annotation. Fortunately, our analysis shows that despite the increased automation the overall TIS annotation quality has increased over the years.

We have developed an iterative PCA-based strategy to evaluate existing TIS annotations. The strength of the strategy is that it employs self-organization and is thus independent of reference data or *a priori* calculations. We have compared between PCA-based and Prodigal TIS annotations for the reference genome of *E. coli*. The analysis showed that both methods supplement each other and that an agreement between the methods is a strong indicator of a correct TIS annotation. Importantly, the addition of the PCA-based strategy to score potential TISs can also be used to 'flag' particular annotations for manual curation. Currently, the iterative PCA-based procedure only uses the positions on PCA component I to score TISs. Integrating scores based on specified features such as RBS sequence, coding/non-coding biases could potentially further improve the accuracy.

Materials and methods

Genome sequences, annotations and sequencing meta-data

Genome sequence and annotation information of all bacterial and archaeal genomes was obtained from the FTP server of NCBI RefSeq (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/) (Pruitt et al., 2012). Genomes with less than 500 ORFs were excluded. The NCBI BioProject database was used to retrieve metadata on the sequencing projects, such as year of sequencing (Barrett et al., 2012). For all species with a sequenced genome that was published before October 2009, additional metadata such as sequencing center, were derived from the GOLD database (Liolios et al., 2010). The SD index, that is the fraction of genes preceded by a Shine-Dalgarno sequence, for a selected set of 277 bacterial and archaeal genomes was taken from (Nakagawa et al., 2010). Taxonomic classifications were retrieved from the NCBI taxonomy database (ftp://ftp.ncbi.nih.gov/pub/taxonomy).

Scripting and data analysis

All automatic procedures were written and executed in Python, whereas the Principal Component Analysis (PCA) and additional statistical analyses were written and executed in R3.0.0 (R Development Core Team, 2008). The distribution plots of alternative TISs were generated using Matplotlib (Hunter, 2007).

Identification of alternative TISs in the context of prior annotation

The genomic position of the annotated translation initiation sites (TISs) for all genes within the microbial RefSeq genomes with >500 annotated open reading frames (ORFs) were taken from the .PTT files available at NCBI. The corresponding coding DNA sequences and upstream regions were collected. Alternative TISs were identified using the nucleotide triplets 'AUG', 'GUG'and 'UUG' as potential start sites and the nucleotide triplets 'UAA', 'UAG' and 'UGA' as stop codons. To identify potential alternative TISs the following criteria were applied: (i) the alternative TIS was in-frame with the annotated TIS; (ii) there was no in-frame stop codon located between the candidate TIS and the annotated TIS; and (iii) the alternative TIS was either found upstream or maximally 198 nucleotides downstream of the annotated TIS. For every genome the distribution of the genomic positions of the alternative TISs with respect to the annotated TISs was calculated. Simultaneously, an expectation of the distribution was calculated based on the sequence properties of the particular genome at hand. The number of alternative TISs in a window

of 198 nucleotides upstream of the 3' end of all annotated ORFs was used to calculate an expected frequency for the occurrence of alternative start codons in the coding part of any gene in the particular genome. A window of 198 nucleotides was chosen because for all studied genomes at this distance the expected total number of alternative in frame TISs has decreased below a total of 10. The expected start codon frequency of occurrence upstream of an annotated TIS (denoted as $f^{start}(obs\_upstream)$) was considered constant and was determined on basis of the average number of observed in-frame alternative starts (independent of in-frame stops) in a window of 198 nucleotides upstream of the longest possible ORFs (independent of the annotated TIS). The expected stop codon frequency of occurrence upstream of an annotated TIS was taken as 3 codons (UAA, UGA and UAG) in 64 and corrected for the AT and GC content of a genome (see *formula 1*). Using the frequencies derived above, we calculated the expected number of alternative TISs (n^{TIS}) upstream at codon position *i* with respect to the annotated TIS, as described in *formula 2* (where N is total number of annotated ORFs):

formula 1:

$$f^{stop}(calc\_upstream) = ([\text{fraction(GC)} + \text{fraction(AT)}/2] * \text{fraction(AT)}^2)/4$$

formula 2:

$$n^{TIS}(i) = N * f^{start}(obs\_upstream) * (1 - f^{stop}(calc\_upstream))$$

The similarity between the distribution of alternative starts -derived using the provided annotation- and the expected distribution of alternative starts -calculated on the basis of the genome sequence- was quantified using a Spearman's rank correlation coefficient.

Principal Component Analysis procedure to assess TISs

An iterative procedure of ten subsequent rounds of PCA was implemented to distinguish TISs on basis of common sequence patterns (procedure depicted in Fig. S4). For each identified candidate TIS in a genome a fixed number of nucleotides upstream and downstream of the annotated start codon were extracted from the sequence file. The upstream and downstream nucleotide sequences were fused to the first nucleotide of the corresponding start codon. The resulting nucleotide sequence was converted to a binary vector in which each position within the sequence was represented by four binary values corresponding to the four different bases (e.g., CTT was thus expressed as 0010 0100 0100). For each annotated ORF the TISs that resulted in the three longest ORFs were used as input for the initial round of PCA. We assumed such a selection included a sufficient number of true TISs to direct the initial PCA. After each round, all alternative and annotated TISs were projected on

the linear combination of the first principle component (PC1) and the PC1-score for each candidate start was computed. The three top scoring candidate TISs for each ORF were then included in the next round of PCA. The number of iterations was set to 10, which was sufficient to converge the PCA results for all genomes tested.

Acknowledgments

This work was part of the BioRange programme of the Netherlands Bioinformatics Centre (NBIC), which is supported by the Netherlands Genomics Initiative (NGI). We thank Martijn van der Pol for his contribution to the project.

Supplementary information

Supplementary data is freely available online:
https://figshare.com/s/ae087404f83cd09b61d8
Python code is made available at:
https://github.com/lexovermars/TISAnalysis

References

Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., Formsma, K., Gerdes, S., Glass, E.M., Kubal, M., et al. (2008). The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9**, 75.

Bakke, P., Carney, N., Deloache, W., Gearing, M., Ingvorsen, K., Lotz, M., McNair, J., Penumetcha, P., Simpson, S., Voss, L., et al. (2009). Evaluation of three automated genome annotations for *Halorhabdus utahensis*. *PloS One* **4**, e6291.

Barrett, T., Clark, K., Gevorgyan, R., Gorelenkov, V., Gribov, E., Karsch-Mizrachi, I., Kimelman, M., Pruitt, K.D., Resenchuk, S., Tatusova, T., et al. (2012). BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acids Res.* **40**, D57-63.

Chang, B., Halgamuge, S., and Tang, S.-L. (2006). Analysis of SD sequences in completed microbial genomes: non-SD-led genes are as common as SD-led genes. *Gene* **373**, 90–99.

Delcher, A.L., Bratke, K.A., Powers, E.C., and Salzberg, S.L. (2007). Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinforma. Oxf. Engl.* **23**, 673–679.

Dunbar, J., Cohn, J.D., and Wall, M.E. (2011). Consistency of gene starts among Burkholderia genomes. *BMC Genomics* **12**, 125.

Ederveen, T.H.A., Overmars, L., and van Hijum, S.A.F.T. (2013). Reduce manual curation by combining gene predictions from multiple annotation engines, a case study of start codon prediction. *PloS One* **8**, e63523.

Hu, G.-Q., Zheng, X., Yang, Y.-F., Ortet, P., She, Z.-S., and Zhu, H. (2008a). ProTISA: a comprehensive resource for translation initiation site annotation in prokaryotic genomes. *Nucleic Acids Res.* **36**, D114-119.

Hu, G.-Q., Zheng, X., Ju, L.-N., Zhu, H., and She, Z.-S. (2008b). Computational evaluation of TIS annotation for prokaryotic genomes. *BMC Bioinformatics* **9**, 160.

Hunter, J.D. (2007). Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **9**, 90–95.

Hyatt, D., Chen, G.-L., Locascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119.

Klassen, J.L., and Currie, C.R. (2013). ORFcor: identifying and accommodating ORF prediction inconsistencies for phylogenetic analysis. *PloS One* **8**, e58387.

Komarova, A.V., Tchufistova, L.S., Dreyfus, M., and Boni, I.V. (2005). AU-rich sequences within 5' untranslated leaders enhance translation and stabilize mRNA in *Escherichia coli. J. Bacteriol.* **187**, 1344–1349.

Krug, K., Carpy, A., Behrends, G., Matic, K., Soares, N.C., and Macek, B. (2013). Deep coverage of the *Escherichia coli* proteome enables the assessment of false discovery rates in simple proteogenomic experiments. *Mol. Cell. Proteomics* MCP **12**, 3420–3430.

Larsen, T.S., and Krogh, A. (2003). EasyGene--a prokaryotic gene finder that ranks ORFs by statistical significance. *BMC Bioinformatics* **4**, 21.

Liolios, K., Chen, I.-M.A., Mavromatis, K., Tavernarakis, N., Hugenholtz, P., Markowitz, V.M., and Kyrpides, N.C. (2010). The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.* **38**, D346-354.

Lukashin, A.V., and Borodovsky, M. (1998). GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.* **26**, 1107–1115.

Ma, J., Campbell, A., and Karlin, S. (2002). Correlations between Shine-Dalgarno Sequences and Gene Features Such as Predicted Expression Levels and Operon Structures. *J. Bacteriol.* **184**, 5733–5745.

Markowitz, V.M., Korzeniewski, F., Palaniappan, K., Szeto, E., Werner, G., Padki, A., Zhao, X., Dubchak, I., Hugenholtz, P., Anderson, I., et al. (2006). The integrated microbial genomes (IMG) system. *Nucleic Acids Res.* **34**, D344-348.

Moll, I., Grill, S., Gualerzi, C.O., and Bläsi, U. (2002). Leaderless mRNAs in bacteria: surprises in ribosomal recruitment and translational control. *Mol. Microbiol.* **43**, 239–246.

Nakagawa, S., Niimura, Y., Miura, K., and Gojobori, T. (2010). Dynamic evolution of translation initiation mechanisms in prokaryotes. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 6382–6387.

Pallejà, A., García-Vallvé, S., and Romeu, A. (2009). Adaptation of the short intergenic spacers between co-directional genes to the Shine-Dalgarno motif among prokaryote genomes. *BMC Genomics* **10**, 537.

Pati, A., Ivanova, N.N., Mikhailova, N., Ovchinnikova, G., Hooper, S.D., Lykidis, A., and Kyrpides, N.C. (2010). GenePRIMP: a gene prediction improvement pipeline for prokaryotic genomes. *Nat. Methods* **7**, 455–457.

Pettis, G.S., Brickman, T.J., and McIntosh, M.A. (1988). Transcriptional mapping and nucleotide sequence of the *Escherichia coli* fepA-fes enterobactin region. Identification of a unique iron-regulated bidirectional promoter. *J. Biol. Chem.* **263**, 18857–18863.

Poptsova, M.S., and Gogarten, J.P. (2010). Using comparative genome analysis to identify problems in annotated microbial genomes. *Microbiology* **156**, 1909–1917.

Pruitt, K.D., Tatusova, T., Brown, G.R., and Maglott, D.R. (2012). NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.* **40**, D130-135.

R Development Core Team (2008). R: A language and environment for statistical computing (R Foundation for Statistical Computing, Vienna, Austria).

Richardson, E.J., and Watson, M. (2013). The automatic annotation of bacterial genomes. *Brief. Bioinform. 14*, 1–12.

Shah, S.P., McVicker, G.P., Mackworth, A.K., Rogic, S., and Ouellette, B.F.F. (2003). GeneComber: combining outputs of gene prediction programs for improved results. *Bioinforma. Oxf. Engl. 19*, 1296–1297.

Shine, J., and Dalgarno, L. (1974). The 3'-terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proc. Natl. Acad. Sci. U. S. A. 71*, 1342–1346.

Skorski, P., Leroy, P., Fayet, O., Dreyfus, M., and Hermann-Le Denmat, S. (2006). The highly efficient translation initiation region from the *Escherichia coli* rpsA gene lacks a shine-dalgarno element. *J. Bacteriol.* **188**, 6277–6285.

Smollett, K.L., Fivian-Hughes, A.S., Smith, J.E., Chang, A., Rao, T., and Davis, E.O. (2009). Experimental determination of translational start sites resolves uncertainties in genomic open reading frame predictions - application to *Mycobacterium tuberculosis*. *Microbiol. Read. Engl.* **155**, 186–197.

Tech, M., Pfeifer, N., Morgenstern, B., and Meinicke, P. (2005). TICO: a tool for improving predictions of prokaryotic translation initiation sites. *Bioinforma. Oxf. Engl.* **21**, 3568–3569.

Udagawa, T., Shimizu, Y., and Ueda, T. (2004). Evidence for the translation initiation of leaderless mRNAs by the intact 70 S ribosome without its dissociation into subunits in *eubacteria*. *J. Biol. Chem.* **279**, 8539–8546.

Wall, M.E., Raghavan, S., Cohn, J.D., and Dunbar, J. (2011). Genome majority vote improves gene predictions. *PLoS Comput. Biol.* **7**, e1002284.

Wang, C., Zhang, M.Q., and Zhang, Z. (2013). Computational identification of active enhancers in model organisms. *Genomics Proteomics Bioinformatics* **11**, 142–150.

Yada, T., Takagi, T., Totoki, Y., Sakaki, Y., and Takaeda, Y. (2003). DIGIT: a novel gene finding program by combining gene-finders. *Pac. Symp. Biocomput. Pac. Symp. Biocomput.* 375–387.

Yok, N.G., and Rosen, G.L. (2011). Combining gene prediction methods to improve metagenomic gene annotation. *BMC Bioinformatics* **12**, 20.

Zheng, X., Hu, G.-Q., She, Z.-S., and Zhu, H. (2011). Leaderless genes in bacteria: clue to the evolution of translation initiation mechanisms in prokaryotes. *BMC Genomics* **12**, 361.

Zhou, J., and Rudd, K.E. (2013). EcoGene 3.0. *Nucleic Acids Res.* **41**, D613-624.

Zhu, H., Hu, G.-Q., Yang, Y.-F., Wang, J., and She, Z.-S. (2007). MED: a new non-supervised gene prediction algorithm for bacterial and archaeal genomes. *BMC Bioinformatics* **8**, 97.

Zur, H., and Tuller, T. (2013). New universal rules of eukaryotic translation initiation fidelity. *PLoS Comput. Biol.* **9**, e1003136.

Chapter 6

Summarizing discussion

Scope of this thesis

The prokaryotic genome sequence is densely packed with functional elements. As a matter of fact, only very little genomic space is left 'unused' (Rogozin et al., 2002). Besides protein coding sequences, which occupy the majority of the genomic space, many other genomic elements with diverse but essential regulatory functions are present. The work presented in this thesis relates to those non-protein coding elements on the DNA. The non-coding elements are involved in the regulation of cellular processes at different 'hierarchical' levels affecting either replication, recombination, transcription or translation. We aimed to contribute and provide depth to the study of the composition and position of the sequence elements with respect to neighboring elements. Within the thesis we did not consider well-studied regulatory elements such as transcription factor binding sites and promoters, but focused on the often unexplored regulatory role of other sequences. This chapter summarizes the main conclusions of this thesis and offers future perspectives.

The analysis and identification of sequence elements is facilitated by integrative visualization of genomic data

Manual (genome-) annotation requires a lot of effort and is time-consuming. Therefore, it is usually beyond the scope of most sequencing projects. As a consequence, most genomes are solely annotated by automated annotation pipelines such as RAST (Aziz et al., 2008), Prokka (Seemann, 2014) and NCBI Prokaryotic Genome Annotation pipeline (Tatusova et al., 2016). These pipelines integrate multiple types of evidence but the quality of function annotation remains dependent on resource-intensive manual curation (Pfeiffer and Oesterhelt, 2015). Moreover, these pipelines are developed for the identification and function annotation of protein- and rRNA/tRNA- coding sequences and do not identify the presence of other sequence elements such as promoters, transcriptional terminators and transcription factors.

The annotation of sequence elements requires the integration of function information on different hierarchical levels. For example, the relative location of sequence elements with respect to the local gene context for regulatory elements is a distinctive feature that relates directly to their function. In the case of transcription factor binding elements the position of the element with respect to the promoter determines the effect of transcription factor binding and consequently the position can be used to distinguish between true and false candidate sites in a binding site prediction. Likewise the global genomic distribution of insertion sequences is an informative feature when exploring their evolution and their effect on genome plasticity. In this thesis we show

that a visual representation of the relative location of sequence elements aids the analysis of potential function, also in the case of new functionality.

Another factor that contributes considerably to the annotation of (unknown-) sequence elements is the inclusion of function information related to the associated genes. Various function classification schemes and methods have been applied to capture the molecular properties and biological role of nucleotide and protein sequences. Some function classification systems, such as COG, are predominantly useful to create a generalized and rapid summary of functionality. Other classification systems, such as PFAM, are more efficient to describe poorly understood function or to enhance existing function annotation. Metabolic annotation systems such as KEGG provide yet another functional context. Moreover, quantitative data such as gene expression (i.e. as derived from RNA-seq and microarrays) can also be valuable to the analysis of the potential function of associated genomic elements. Gene expression data can, for example, be used to infer a regulatory network, or to evaluate an already reconstructed network.

In **chapter 2**, we describe two user-friendly visualization tools that were developed to aid the annotation and analysis of genomic elements. Both tools provide a visual access to publicly available genomic and annotation data. The latter include protein names, sequence characteristics, PFAM domains (Punta et al., 2012), subcellular location predictions (Yu et al., 2010) and COG categories (Tatusov et al., 2003). Importantly, both tools allow the user to include predicted positions of (regulatory-) elements or data from RNA-seq and microarray experiments, and thereby allow easy integration of in-house data with data available in the public domain.

The Microbial Genomic context Viewer (MGcV; http://mgcv.cmbi.ru.nl), described in **chapter 2a**, is an interactive visualization tool tailored to facilitate small-scale genome analysis of publicly available genomes (last update: March 2015). It renders a visualization of the genomic context of any set of selected genes, genes within a phylogenetic tree, genomic segments, or regulatory elements. Its interactive gene context maps allow users to graphically select sets of genes and export data for subsequent analysis. The various input types and practical data export functionality make MGcV a unique tool to facilitate small-scale genome analysis such as annotation of gene function, discovery of regulatory elements or the sequence-based reconstruction of regulatory networks. We have used the tool extensively to support our comparative genome analyses (this thesis, **Chapter 4**; Francke et al., 2011, 2011; Khayatt et al., 2013; Liu et al., 2012; Siezen et al., 2012) and find that other groups from all over the world use the tool similarly (Ejby

et al., 2016; van den Esker et al., 2016; Gennaris et al., 2015; Greening et al., 2016; Houdt and Mergeay, 2015; Lagares et al., 2016; Lee et al., 2014; Ney et al., 2016; Rauch et al., 2014; Roche et al., 2015; Sorci et al., 2014; Xiao et al., 2015).

In addition to the visualization of local genomic context, we decided to develop a web-application that allows users to create visually appealing circular genome-wide maps. The Circular genome Viewer (CiVI; http://civi.cmbi.ru.nl), described in **chapter 2b**, is a web-application in which users can create custom circular genome visualizations in a stepwise manner. Similar to MGcV, various types of annotation data are available for each publicly available genome (last update: March 2015). These data can again be integrated with in-house data like, for example, RNA-seq data and the predicted location of genomic elements. The tool offers unique features to study the genomic distribution of predicted genomic elements in greater detail. The features include an automatically generated distance distribution of the elements with respect to the neighboring genes and a distribution summarizing the local organization of these genes. The added features laid a solid foundation for the study of repeated sequence elements with unknown function as reported in **chapters 3** and **4**. Moreover, the tool enables users to conveniently export the annotation of genomic context of uploaded elements. We used CiVi to support our genome analyses (this thesis, **Chapter 3** and **Chapter 4**) and like MGcV, CiVI proves a useful extension of the tools available to the scientific community (Ahlstrom et al., 2016; Choi, 2016; Mehla and Ramana, 2016; Sghaier et al., 2016; Sheibani-Tezerji et al., 2015).

Identification and global characterization of repeated sequences in prokaryotic genomes

Bacteria and archaeal genomes tend to be very compact, with protein coding sequences occupying around 90% of the genome for most of the species. The intergenic regions are usually small but are packed with various elements including sequences that are repeated throughout the chromosome. However, the physiological roles of many of these repeated sequences remain unknown. To uncover new sequences of interest with potential regulatory roles, we decided to investigate the presence of repeated sequences in prokaryotic genomes.

In **chapter 3** we empirically identified overrepresented dodecamers (i.e. sequences of 12 nt) within the intergenic regions of 1516 prokaryotic genomes. We considered that newly identified overrepresented dodecamers could potentially serve as marker sequences to distinguish different bacterial

species. Moreover, we decided to look for straightforward parameters that could indicate the structural or functional role of the repeated sequences within the genomes. We decided to focus on sequences of 12 conserved nt to ensure a high recovery and at the same time to decrease the probability of random occurrences of the identified sequences. Nevertheless, the number of markers could be increased by allowing various target lengths and some motif degeneracy.

To characterize newly identified sequence elements we formulated a generalized strategy in which a distribution profile is created for every repeated sequence, describing: i) the abundance and genome-wide distribution, ii) the taxonomic distribution and iii) the distribution with respect to the local gene organization. **Chapter 3** describes various repeats and illustrates the application of our generalized strategy.

We found 583 overrepresented dodecamer sequences with varying and sometimes intriguing abundance and distribution profiles making them potentially very useful in the identification of specific species or strains. Some of them were described before but for many others we have not found previous reports. Moreover, in almost all cases a notion of their biological role is lacking. The most widely spread and abundant repeated sequences we found were Adenine-rich. These Adenine-rich repeats were highly abundant and uniformly distributed in the genomes in which they were identified and, although they were found mostly within intragenic regions, they did not seem to exhibit a positional bias with respect to local gene organization. Such a distribution profile indicates a global role, but most likely no gene-organization related function, such as gene transcription regulation. At the same time, we found sequences that were highly specific for various species and strains. For example, the repeat sequence ATGCCGTCTGAA appears highly specific for the genus *Neisseria*.

Using a combination of different perspectives on the distribution and abundance of a sequence, i.e. a broad taxonomic view, a genome-wide view and view focused on local gene organization we were able to provide a description of the identified repeated sequences. Moreover, we argue that a similar organization could point at a similar role of different sequences in different species. Our generalized strategy may serve as a foundation for further in-depth analysis of the function of such sequences. The provided distribution profile can be used to select sequences that match the characteristics of sequences for which the function is known. Vice versa, we illustrate that the distribution profile can also be used to narrow down the list of potential functions. We did not assign a detailed function to all retrieved sequences, as

this would have required substantial additional analyses, beyond the scope of this chapter. Instead, we chose a well-known repeat with unknown function found in *E. coli* and were able to elucidate a potential function using the developed strategy as described in **chapter 4**.

Repetitive Extragenic Palindromic elements have a common topological role in the reduction of transcriptional interference

One of the repeat sequences with an intriguing distribution and abundance profile identified in **chapter 3** are commonly denoted as Repetitive Extragenic Palindromic elements (REPs). REPs are short palindromic sequences that can be very abundant in some enteric bacteria, where they occupy up to 1% of the genome sequence. Various roles have been attributed to REPs in the past, but none of them have provided a common functional denominator.

In **chapter 4** we aimed to uncover the primary role of REPs by analyzing their positional distribution in the genome of *E. coli* K12 MG1655 and analyzing the sequence characteristics and expression dynamics of the neighboring genes. Within the genome of *E. coli* K12 MG1655, REPs are found exclusively between co-oriented and convergent gene-pairs. We have shown that the REPs are mostly located close to the 3' end of the adjacent genes. We analyzed the sequence characteristics of the adjacent genes and found that the upstream members of co-oriented gene-pairs (i.e. '→' REP →) have a Codon Adaptation Index (CAI) that is significantly higher than their downstream members (i.e. → REP '→'). The CAI values of convergent genes adjacent to a REP element are also significantly higher than their counterparts without (i.e. '→ REP ←' vs. '→ ←'). These distinct differences were also apparent when we used actual gene expression data from 466 microarray experiments (Faith et al., 2008). Our analysis on this dataset indicated that the average difference in expression between convergent REP-associated gene-pairs was lower compared to non-REP-associated gene-pairs for every experiment in the data set, even though their average expression was higher. This observation and the fact that REPs are located at the 3' end of highly expressed genes, and positioned in gene organizations prone to transcription induced supercoiling, lead us to a new hypothesis concerning the role of REPs, namely: that REPs reduce the negative effects of transcription induced supercoiling. REPs may relieve the supercoiling stress by the formation of cruciform structures, allowing an overall higher expression of convergent gene-pairs and of genes located downstream of highly expressed genes. Given the spread of REP-like elements in the species of the *gammaproteobacteria* one may assume that this mechanism should also be active in other species than the enteric bacteria.

In this thesis, we thus have linked the well-known REP elements of *E. coli* for the first time to a consistent putative biological role, namely: the reduction of transcription interference caused by transcription induced supercoiling. Further experimental studies could elucidate this important but unexplored role of REPs. One experimental approach would be to make use of supercoiling sensitive promoters to test the impact of REP sequences for different gene organizations. Measuring supercoiling levels *in vivo* would provide a significant advance in characterizing the role of REPs and supercoiling stress in transcription regulation. However, measuring the supercoiling state *in vivo* is still challenging, although novel techniques such as the use of magnetic tweezers combined with fluorescence seem promising (Vlijm et al., 2015). Supercoiling in a prokaryotic cell is dynamic and is responsive to changes in the external environment of the cell (Dorman, 2011). Changes in growth phase, pH, temperature, osmotic pressure and carbon sources have already been shown to result in changes in supercoiling state of bacterial DNA (Cameron et al., 2011). It was noted before that bacteria exploit supercoiling effects as part of their transcriptional regulatory machinery (Cameron et al., 2011; Cheung et al., 2003; Dorman, 1991), and it was established that it can influence gene expression both on a global and a local level (Blot et al., 2006; Ferrándiz et al., 2014). It is therefore tempting to speculate that DNA supercoiling is a key factor in the regulatory network of *E. coli*, and in other organisms. The impact on supercoiling of specific sequence elements adds another layer to the complexity of the regulatory network of prokaryotic transcription.

Two new methods to evaluate the quality of translation initiation site annotations.

The *ab initio* identification of coding sequences is not only the first step in the annotation of a genome, it is also an essential step for many subsequent analyses (Bauer et al., 2010). Various types of errors can be introduced during (computational) gene-calling (Dunbar et al., 2011). True coding regions can be missed and predicted coding regions might not represent a true coding region. However, the most challenging aspect of the identification of coding sequences is the identification of the correct translation initiation site (TIS). Incorrect TIS annotations can impair the identification of regulatory elements and N-terminal signal peptides and can misguide sequence alignments. Various computational methods, such as Prodigal (Hyatt et al., 2010) and Glimmer (Delcher et al., 2007), are generally used to identify coding sequences with a (relatively) low error rate. Yet, the identification of genes in high GC- and low GC-content genomes remains challenging. Moreover, the available public resources contain a substantial number of annotations of less quality.

In order to address this problem, we first formulated (**first part of chapter 5)** a reference-free method to assess the TIS annotation quality of a genome. We found that the correlation between a) the observed distribution of all potential TISs in the context of all predicted coding regions and b) the expected distribution of potential TISs, provides a straightforward quality measure. We validated the approach by scoring the TIS annotation quality of various model organisms with high quality annotations. Our method then enabled us to evaluate the TIS annotation quality of all available bacterial and archaeal Refseq genomes. We established that the TIS annotation of 87% of the evaluated public genomes is of appropriate quality, but also that approximately 13% of the public annotations requires a significant improvement. The dataset also allowed us to analyze various factors that are believed to affect TIS annotation quality. Interestingly, out of the factors that were taken in to account, only the GC-content of the genomes correlated clearly with TIS annotation quality. Fortunately it seems that the increasing representation of automated annotations does not result in a decrease in annotation quality throughout the years.

In the **second part of chapter 5** we describe a strategy to evaluate and -where possible- improve existing TIS annotations. This strategy employs an iterative principal component analysis (PCA) and is independent of any reference data or *a priori* calculations. We applied a simple scoring scheme to utilize the scores from different iterations of PCA and therewith re-annotated the TISs for various genomes. We found that many of the badly annotated genomes within a data set of 277 genomes could be improved when we applied this simple scoring scheme. We compared the quality of the PCA-adjusted TIS annotations with the ones derived from a re-annotation using Prodigal and found that both methods improve the annotations in various genomes. Our PCA-based strategy can supplement existing computational approaches as it can be used to evaluate their outcome and 'flag' particular annotations that might require manual curation.

Future perspectives

In this thesis, we have described various efforts to improve and facilitate the analysis of the function of genomic elements. We have implemented two visualization tools that provide an integrated, visual access to public genome data, and that facilitate in depth as well as genome-wide analyses of sequence elements. One feature that is commonly requested by users is the ability to upload and include non-public genomes. From a user's perspective, the ideal solution might include annotation of any uploaded genome sequence. However, as genome annotation is a relatively resource intensive task, it might

be sustainable to enable users to import annotated genomes from annotation platforms such as RAST (Aziz et al., 2008) and Prokka (Seemann, 2014) and supplement those with additional annotation. Currently, a relatively limited set of function descriptors is available within our visualization tools and these are not connected to network representations as already available in some resources. For instance, gene products could be connected in functional networks such as metabolic pathways (e.g. KEGG (Kanehisa et al., 2015), protein complexes (Caufield et al., 2015), or protein interaction networks (e.g. STRING (Franceschini et al., 2013)). Addition of network representations could be valuable when studying genomic elements as they provide additional functional context to the genes. Unfortunately, the quality of automated *in silico* approaches for the reconstruction of metabolic networks and pathways is still limited (Hamilton and Reed, 2014). This is most notable for metabolic pathways that are only present in non-model organisms and non-essential metabolic pathways. Nevertheless, efforts in improving the quality of these approaches are ongoing and methods to create generic 'metabolic network visualizations' from metabolic annotation data have already been developed. Considering these ongoing efforts, a visualization tool that provides a simultaneous and fully integrated view of i) local genomic context, ii) genome-wide context (i.e. genome-wide distribution and genome-wide trends) and iii) network context (i.e. metabolic context and/or protein interactions) will become feasible in the future.

A visualization tool with simultaneous views on multiple hierarchical levels of genomic data, could be considered a multi-dimensional visualization of genome information. As biology is complex and most genome elements are involved in various hierarchical levels within a cell, multi-dimensional genome annotation is important for describing and capturing the cell's functional capabilities. Therefore, a truly multi-dimensional visualization could be a valuable improvement over the tools presented in this thesis and advance, for example, the evaluation and refinement of reconstructed regulatory networks. Moreover, it could facilitate researchers to assess their findings in various, (unexplored-) contexts and thereby promote new findings.

We have identified a large number of overrepresented dodecamers within the intergenic regions of 1516 prokaryotic genomes. These newly identified dodecamers could potentially serve as marker sequences to distinguish different bacterial species. Amplification of short genomic fragments lying between repetitive elements by using low-stringency PCR with outward primers, denoted as rep-PCR, is an efficient method for bacterial typing. We have shown example sequences with very specific abundance profiles that are useful to distinguish different species and even strains. The identified

sequences and their taxonomic distribution and abundance profiles could be made available to the scientific community in an interactive database. In such a database, a user would be able to select the bacterial species or strains that one would like to distinguish and a list with potential sequences and their profiles would be returned. Such a platform could be expanded with *in silico* PCR amplification within selected genomes and thereby provide a reference for the applicability of potential marker sequences for molecular typing using, for example, rep-PCR. Major advantages of rep-PCR approaches include flexibility, technical simplicity and rapidity (Struelens, 1996). However, rep-PCR patterns are difficult to interpret and non-exchangeable between laboratories due to variable fragment amplification efficiency. The recent developments in next generation sequencing techniques such as SLST are enabling higher resolution (Scholz et al., 2014).

The identification of translation initiation sites (TISs) constitutes an important aspect of sequence-based genome analysis. It would therefore be beneficial when researchers assess the TIS annotation quality of genomes that they include in their (comparative) genome analyses. Genomes with a poor TIS annotation quality should either be re-annotated or discarded from the data set. We assessed all available complete prokaryotic genomes in the Refseq database and deposited their TIS annotation quality scores. It would be valuable when the scientific community could assess the TIS annotation quality of any genome using for example a web-application. The presented method in **chapter 5** does not require any reference data and does not require extensive resources or computation time, thus providing a good starting point for such an application. Similarly, the PCA-based TIS re-annotation method we formulated could be generally implemented. Currently, the method works remarkably well given the fact that the prediction relies solely on one parameter, the score related to the first component in the PCA. Integrating different scores based on different, specific genome features such as ribosomal binding site and coding/non-coding sequence bias could therefore improve the accuracy. Yet, the lack of experimentally verified data sets makes it more challenging to improve *in silico* methods. Efforts to create such sets and improved assays are underway. For instance, Smollett and colleagues devised a new assay using a combination of epitope tagging and frameshift mutagenesis (Smollett et al., 2009).

Concluding remarks

In conclusion, the research reported in this thesis focused on non-coding sequence elements with a (potential-) regulatory role. We have facilitated the bioinformatics analysis of these sequence elements by creating visualization

tools with specifically tailored functionality and by assessing and improving the annotation of transcription initiation sites. Our strategy to characterize sequence elements based on location and taxonomy can be of great help to identify and select sequences with a potential regulatory role. The analysis of the REP elements described in **chapter 4** illustrates that the analyses of uncharacterized sequence elements can be considerably strengthened through a high quality of the annotation of coding regions and integrative genome visualization. Identification of the functional role of recurring sequence elements, in for example transcription regulation or genome organization, will help us to further our understanding of the complex path from genotype to phenotype.

References

Ahlstrom, C., Barkema, H.W., and De Buck, J. (2016). Relative frequency of 4 major strain types of *Mycobacterium avium* ssp. paratuberculosis in Canadian dairy herds using a novel single nucleotide polymorphism-based polymerase chain reaction. *J. Dairy Sci.* **99**, 8297–8303.

Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., Formsma, K., Gerdes, S., Glass, E.M., Kubal, M., et al. (2008). The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9**, 75.

Bauer, A.L., Hlavacek, W.S., Unkefer, P.J., and Mu, F. (2010). Using Sequence-Specific Chemical and Structural Properties of DNA to Predict Transcription Factor Binding Sites. *PLoS Comput Biol* **6**, e1001007.

Blot, N., Mavathur, R., Geertz, M., Travers, A., and Muskhelishvili, G. (2006). Homeostatic regulation of supercoiling sensitivity coordinates transcription of the bacterial genome. *EMBO Rep.* **7**, 710–715.

Cameron, A.D.S., Stoebel, D.M., and Dorman, C.J. (2011). DNA supercoiling is differentially regulated by environmental factors and FIS in *Escherichia coli* and *Salmonella enterica. Mol. Microbiol.* **80**, 85–101.

Caufield, J.H., Abreu, M., Wimble, C., and Uetz, P. (2015). Protein Complexes in Bacteria. *PLOS Comput Biol* **11**, e1004107.

Cheung, K.J., Badarinarayana, V., Selinger, D.W., Janse, D., and Church, G.M. (2003). A microarray-based antibiotic screen identifies a regulatory role for supercoiling in the osmotic stress response of *Escherichia coli*. *Genome Res.* **13**, 206–215.

Choi, S.C. (2016). On the study of microbial transcriptomes using second- and third-generation sequencing technologies. *J. Microbiol. Seoul Korea* **54**, 527–536.

Delcher, A.L., Bratke, K.A., Powers, E.C., and Salzberg, S.L. (2007). Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* **23**, 673–679.

Dorman, C.J. (1991). DNA supercoiling and environmental regulation of gene expression in pathogenic bacteria. *Infect. Immun.* **59**, 745–749.

Dorman, C.J. (2011). Regulation of transcription by DNA supercoiling in *Mycoplasma genitalium*: global control in the smallest known self-replicating genome. *Mol. Microbiol.* **81**, 302–304.

Dunbar, J., Cohn, J.D., and Wall, M.E. (2011). Consistency of gene starts among *Burkholderia* genomes. *BMC Genomics* **12**, 125.

Ejby, M., Fredslund, F., Andersen, J.M., Vujičić Žagar, A., Henriksen, J.R., Andersen, T.L., Svensson, B., Slotboom, D.J., and Abou Hachem, M. (2016). An ATP Binding Cassette Transporter Mediates the Uptake of α-(1,6)-Linked Dietary Oligosaccharides in *Bifidobacterium* and Correlates with Competitive Growth on These Substrates. *J. Biol. Chem.* **291**, 20220–20231.

van den Esker, M.H., Kovács, Á.T., and Kuipers, O.P. (2016). YsbA and LytST are essential for pyruvate utilization in *Bacillus subtilis*. *Environ. Microbiol*.

Faith, J.J., Driscoll, M.E., Fusaro, V.A., Cosgrove, E.J., Hayete, B., Juhn, F.S., Schneider, S.J., and Gardner, T.S. (2008). Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata. *Nucleic Acids Res.* **36**, D866-870.

Ferrándiz, M.-J., Arnanz, C., Martín-Galiano, A.J., Rodríguez-Martín, C., and Campa, A.G. de la (2014). Role of Global and Local Topology in the Regulation of Gene Expression in *Streptococcus pneumoniae*. *PLOS ONE* **9**, e101574.

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., Mering, C. von, et al. (2013). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* **41**, D808–D815.

Francke, C., Groot Kormelink, T., Hagemeijer, Y., Overmars, L., Sluijter, V., Moezelaar, R., and Siezen, R.J. (2011). Comparative analyses imply that the enigmatic sigma factor 54 is a central controller of the bacterial exterior. *BMC Genomics* **12**, 385.

Gennaris, A., Ezraty, B., Henry, C., Agrebi, R., Vergnes, A., Oheix, E., Bos, J., Leverrier, P., Espinosa, L., Szewczyk, J., et al. (2015). Repairing oxidized proteins in the bacterial envelope using respiratory chain electrons. *Nature* **528**, 409–412.

Greening, C., Biswas, A., Carere, C.R., Jackson, C.J., Taylor, M.C., Stott, M.B., Cook, G.M., and Morales, S.E. (2016). Genomic and metagenomic surveys of hydrogenase distribution indicate H2 is a widely utilised energy source for microbial growth and survival. *ISME J.* **10**, 761–777.

Hamilton, J.J., and Reed, J.L. (2014). Software platforms to facilitate reconstructing genome-scale metabolic networks. *Environ. Microbiol.* **16**, 49–59.

Houdt, R.V., and Mergeay, M. (2015). Genomic Context of Metal Response Genes in *Cupriavidus metallidurans* with a Focus on Strain CH34. In Metal Response in *Cupriavidus Metallidurans*, M. Mergeay, and R.V. Houdt, *eds. (Springer International Publishing)*, pp. 21–44.

Hyatt, D., Chen, G.-L., LoCascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2015). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* gkv1070.

Khayatt, B.I., Overmars, L., Siezen, R.J., and Francke, C. (2013). Classification of the adenylation and acyl-transferase activity of NRPS and PKS systems using ensembles of substrate specific hidden Markov models. *PloS One* **8**, e62136.

Lagares, A., Roux, I., and Valverde, C. (2016). Phylogenetic distribution and evolutionary pattern of an α-proteobacterial small RNA gene that controls polyhydroxybutyrate accumulation in Sinorhizobium meliloti. *Mol. Phylogenet. Evol. 99*, 182–193.

Lee, I.-C., van Swam, I.I., Tomita, S., Morsomme, P., Rolain, T., Hols, P., Kleerebezem, M., and Bron, P.A. (2014). GtfA and GtfB are both required for protein O-glycosylation in *Lactobacillus plantarum*. *J. Bacteriol.* **196**, 1671–1682.

Liu, M., Prakash, C., Nauta, A., Siezen, R.J., and Francke, C. (2012). A computational analysis of cysteine and methionine metabolism and its regulation in dairy starter and related bacteria. *J. Bacteriol.* JB.06816-11.

Mehla, K., and Ramana, J. (2016). Identification of epitope-based peptide vaccine candidates against enterotoxigenic *Escherichia coli*: a comparative genomics and immunoinformatics approach. *Mol. Biosyst.* **12**, 890–901.

Ney, B., Ahmed, F.H., Carere, C.R., Biswas, A., Warden, A.C., Morales, S.E., Pandey, G., Watt, S.J., Oakeshott, J.G., Taylor, M.C., et al. (2016). The methanogenic redox cofactor F420 is widely synthesized by aerobic soil bacteria. *ISME J.*

Pfeiffer, F., and Oesterhelt, D. (2015). A Manual Curation Strategy to Improve Genome Annotation: Application to a Set of Haloarchael Genomes. **Life** *5*, 1427–1444.

Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., et al. (2012). The Pfam protein families database. *Nucleic Acids Res.* **40**, D290-301.

Rauch, B.J., Gustafson, A., and Perona, J.J. (2014). Novel proteins for homocysteine biosynthesis in anaerobic microorganisms. *Mol. Microbiol.* **94**, 1330–1342.

Roche, B., Agrebi, R., Huguenot, A., Ollagnier de Choudens, S., Barras, F., and Py, B. (2015). Turning *Escherichia coli* into a Frataxin-Dependent Organism. *PLoS Genet.* **11**, e1005134.

Rogozin, I.B., Makarova, K.S., Natale, D.A., Spiridonov, A.N., Tatusov, R.L., Wolf, Y.I., Yin, J., and Koonin, E.V. (2002). Congruent evolution of different classes of non-coding DNA in prokaryotic genomes. *Nucleic Acids Res.* **30**, 4264–4271.

Scholz, C.F.P., Jensen, A., Lomholt, H.B., Brüggemann, H., and Kilian, M. (2014). A Novel High-Resolution Single Locus Sequence Typing Scheme for Mixed Populations of *Propionibacterium acnes* In Vivo. *PLOS ONE* **9**, e104199.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinforma. Oxf. Engl.* **30**, 2068–2069.

Sghaier, H., Hezbri, K., Ghodhbane-Gtari, F., Pujic, P., Sen, A., Daffonchio, D., Boudabous, A., Tisa, L.S., Klenk, H.-P., Armengaud, J., et al. (2016). Stone-dwelling *actinobacteria Blastococcus saxobsidens, Modestobacter marinus* and *Geodermatophilus obscurus* proteogenomes. *ISME J.* **10**, 21–29.

Sheibani-Tezerji, R., Rattei, T., Sessitsch, A., Trognitz, F., and Mitter, B. (2015). Transcriptome Profiling of the Endophyte *Burkholderia phytofirmans* PsJN Indicates Sensing of the Plant Environment and Drought Stress. *mBio* **6**, e00621-615.

Siezen, R.J., Francke, C., Renckens, B., Boekhorst, J., Wels, M., Kleerebezem, M., and van Hijum, S.A.F.T. (2012). Complete resequencing and reannotation of the *Lactobacillus plantarum* WCFS1 genome. *J. Bacteriol.* **194**, 195–196.

Smollett, K.L., Fivian-Hughes, A.S., Smith, J.E., Chang, A., Rao, T., and Davis, E.O. (2009). Experimental determination of translational start sites resolves uncertainties in genomic open reading frame predictions - application to *Mycobacterium tuberculosis*. *Microbiol. Read. Engl.* **155**, 186–197.

Sorci, L., Ruggieri, S., and Raffaelli, N. (2014). NAD homeostasis in the bacterial response to DNA/RNA damage. *DNA Repair* **23**, 17–26.

Struelens, M.J. (1996). Consensus guidelines for appropriate use and evaluation of microbial epidemiologic typing systems. *Clin. Microbiol. Infect.* **2**, 2–11.

Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., et al. (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics 4*, 41.

Tatusova, T., DiCuccio, M., Badretdin, A., Chetvernin, V., Nawrocki, E.P., Zaslavsky, L., Lomsadze, A., Pruitt, K.D., Borodovsky, M., and Ostell, J. (2016). NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res.* **44**, 6614–6624.

Vlijm, R., Mashaghi, A., Bernard, S., Modesti, M., and Dekker, C. (2015). Experimental phase diagram of negatively supercoiled DNA measured by magnetic tweezers and fluorescence. *Nanoscale* **7**, 3205–3216.

Xiao, Y., van Hijum, S.A.F.T., Abee, T., and Wells-Bennik, M.H.J. (2015). Genome-Wide Transcriptional Profiling of *Clostridium perfringens* SM101 during Sporulation Extends the Core of Putative Sporulation Genes and Genes Determining Spore Properties and Germination Characteristics. *PloS One* **10**, e0127036.

Yu, N.Y., Wagner, J.R., Laird, M.R., Melli, G., Rey, S., Lo, R., Dao, P., Sahinalp, S.C., Ester, M., Foster, L.J., et al. (2010). PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinforma. Oxf. Engl.* **26**, 1608–1615.

Addenda

Nederlandse Samenvatting

Acknowledgements

About the Author

List of publications

A

ADDENDUM I

Nederlandse samenvatting

Het DNA van bacteriën en archaea bestaat vrijwel uitsluitend uit functionele elementen. Slechts weinig DNA ruimte blijft ongebruikt. De meerderheid van de elementen, de genen, wordt overgeschreven in de vorm van RNA (transcriptie) en daarna vertaald in de vorm van eiwit (translatie). Daarnaast zijn er elementen met die een rol spelen bij de vorming van structuur en/of het recruteren van moleculen. Dit proefschrift beschrijft de analyse van deze niet coderende elementen en van hun betrokkenheid bij de regulatie van belangrijke cellulaire processen zoals replicatie, recombinatie, transcriptie en translatie. In het beschreven onderzoek werd de compositie, positie en context gebruikt om diepte aan te brengen in de studie van deze elementen. De aandacht werd daarbij niet zozeer gericht op reeds goed bestudeerde regulatoire elementen zoals promotoren en DNA-sequenties die herkend worden door transcriptiefactoren, maar juist op de elementen met een nog onbekende regulatoire rol. Voor eiwit coderende sequenties zijn veel wetenschappelijke gereedschappen en strategieën beschikbaar ten behoeve van het achterhalen van de functie. Dit was echter veel minder het geval voor niet-coderende sequenties. Daarom zijn in het kader van dit proefschrift tools ontwikkeld ten behoeve van het faciliteren van het onderzoek naar hun functie.

Visuele en geïntegreerde toegang tot genomische data bevordert de analyse en identificatie van sequentie elementen

Vrijwel alle nieuwe genoom sequencties worden automatisch geannoteerd (d.w.z. genen worden voorzien van een functie) door aan elkaar gekoppelde software. Deze softwareoplossingen integreren gewoonlijk meerdere typen bewijs om tot de beste beschrijving van de functie te komen. Desondanks blijkt de kwaliteit van de uiteindelijke annotatie nog steeds grotendeels afhankelijk van een tijdrovende handmatige inspectie en aanpassing achteraf, de curatie. Het geautomatiseerde annotatieproces betreft de identificatie en functie annotatie van eiwit- en rRNA/tRNA- coderende sequenties en hun start. De identificatie van de andere elementen op het genoom, zoals promoteren, transcriptie terminatoren en transcriptie factoren vindt plaats aan de hand van de herkenning van referentiemotieven.

De annotatie van niet-coderende sequentie-elementen vereist de integratie van functie informatie op verschillende hiërarchische niveaus. Bijvoorbeeld: de relatieve locatie van sequentie-elementen ten opzichte van de plaatselijke

gen context is voor regulerende elementen een onderscheidend kenmerk dat direct betrekking heeft op hun functie. In het geval van transcriptiefactor-bindende elementen is het de positie van het element ten opzichte van de promoter die hun effect bepaalt, waarbij elementen op de promoter of downstream er van de transcriptie belemmeren (repressie), en elementen net upstream van de promoter de transcriptie bevorderen (activatie). Daarmee kan de positie gebruikt worden om onderscheid te maken tussen echte en valse kandidaten in een voorspelling. Op een soortgelijke manier kan de globale verdeling binnen een genoom van insertie sequenties als informatieve eigenschap dienen bij het verkennen van hun evolutie en hun effect op genoom plasticiteit. In dit proefschrift wordt aangetoond dat een visuele representatie van de relatieve locatie van sequentie-elementen bijdraagt aan de analyse van mogelijke functies, ook in het geval van onbekende functionaliteit.

Een andere factor die aanzienlijk kan bijdragen aan de annotatie van (onbekende-) sequentie-elementen is het opnemen van functie-informatie van de omliggende genen. Verschillende functie classificatieschema's en classificatie-methoden kunnen toegepast worden om de moleculaire eigenschappen en biologische rol van de nucleotiden- en eiwitsequenties te omschrijven. Sommige functie-classificatiesystemen, zoals COG, zijn voornamelijk nuttig om een gegeneraliseerde en snelle samenvatting van functionaliteit te creëren. Andere classificatiesystemen, zoals PFAM, zijn geschikter om slecht begrepen functionaliteit te beschrijven of bestaande functie annotatie verbeteren. Annotatiesystemen op metabool niveau, zoals KEGG, geven nog een aanvullend functioneel perspectief in de vorm van geassocieerde stoffen (compounds), reacties en routes. Bovendien kunnen kwantitatieve gegevens zoals gen-expressie (d.w.z. zoals afgeleid van RNA-seq en DNA-microarrays) ook waardevol zijn voor de analyse van de mogelijke functie van genomische elementen. Gen-expressie data kunnen bijvoorbeeld gebruikt worden om een regulatoir netwerk te reconstrueren of een reeds gereconstrueerd netwerk evalueren.

De Microbial Genomic context Viewer (MGcV; http://mgcv.cmbi.ru.nl), beschreven in **hoofdstuk 2a**, is een interactieve applicatie die is toegespitst op het faciliteren van genomische analyses van publiek beschikbare genomen. De web-applicatie genereert een visuele weergave van de genomische context van elk mogelijke set van geselecteerde genen, fylogenetische bomen, segmenten van genomen of elementen met regulatoire functie. De interactieve kaarten maken het mogelijk voor gebruikers om groepen van genen op grafische wijze te selecteren en corresponderende data te exporteren voor opvolgende analyses. De verschillende invoer mogelijkheden en praktische exporteer

functionaliteiten maken MGcV een unieke applicatie om genomische analyses zoals de annotatie van gen functie, het ontdekken van regulatoire elementen of het reconstrueren van regulatoire netwerken te faciliteren. Wij hebben deze applicatie veelvuldig toegepast om onze genoom-vergelijkende analyses (dit proefschrift, **hoofdstuk 4**; Francke et al., 2011, 2011; Khayatt et al., 2013; Liu et al., 2012; Siezen et al., 2012) te ondersteunen en hebben ondervonden dat onderzoeksgroepen over de hele wereld deze applicatie op een soortgelijke manier toepassen (Ejby et al., 2016; van den Esker et al., 2016; Gennaris et al., 2015; Greening et al., 2016; Houdt and Mergeay, 2015; Lagares et al., 2016; Lee et al., 2014; Ney et al., 2016; Rauch et al., 2014; Roche et al., 2015; Sorci et al., 2014; Xiao et al., 2015).

Als toevoeging op de lokale genomische context visualisaties werd een web-applicatie ontwikkeld die het mogelijk maakt om visueel aantrekkelijke circulaire kaarten van complete genomen te creëren. De Circular genome Viewer (CiVI; http://civi.cmbi.ru.nl/), beschreven in **hoofdstuk 2b**, is een webapplicatie waarmee gebruikers persoonlijk aangepaste circulaire kaarten van bacteriële genomen stapsgewijs kunnen creëren. Net als in MGcV zijn verschillende type annotatie data beschikbaar voor elk publiek beschikbaar genoom. Deze data kunnen geïntegreerd worden met eigen data zoals RNA-seq data en voorspelde locaties van genomische elementen. De web-applicatie biedt unieke mogelijkheden om de genomische distributie van voorspelde elementen in detail te bestuderen. Tot deze mogelijkheden behoren bijvoorbeeld de automatisch gegenereerde afstands-distributie van de elementen ten opzichte van in de buurt liggende genen en een distributie waarin de lokale gen-organisatie wordt samengevat. Deze toegevoegde mogelijkheden vormen een belangrijke fundering voor het in **hoofdstuk 3** en hoofdstuk 4 gerapporteerde onderzoek aan herhalende sequentie elementen met onbekende functie. Wij hebben CiVi gebruikt om onze genoom-analyses te ondersteunen (dit proefschrift, **hoofdstuk 3** en **hoofdstuk 4**) en, net als MGcV, blijkt CiVi voor de wetenschappelijke gemeenschap ook een nuttige toevoeging op de beschikbare applicaties (Ahlstrom et al., 2016; Choi, 2016; Mehla and Ramana, 2016; Sghaier et al., 2016; Sheibani-Tezerji et al., 2015).

Identificatie en karakterisering van zich herhalende sequenties in prokaryote genomen

De genomen van bacteriën en archaea zijn overal het algemeen erg compact. In de meeste soorten, bestaan de genomen voor ongeveer 90% uit eiwit coderende sequenties. De regio's tussen deze eiwit coderende sequenties, ook wel bekend als intergene regio's, zijn vaak klein en bevatten allerlei soorten herhaalde sequenties. De functie van veel van deze herhaalde sequenties is

vooralsnog onbekend. Om nieuwe sequenties te ontdekken met een potentieel regulerende rol hebben wij de aanwezigheid van deze herhaalde sequenties in de genomen van prokaryoten onderzocht.

In **hoofdstuk 3,** hebben wij oververtegenwoordigde dodecameren (of te wel sequenties van 12 nucleotiden lang) geïdentificeerd in de intergene regio's van de genomen van 1516 prokaryoten. De gedachte daarbij was dat nieuw geïdentificeerde sequenties gebruikt kunnen worden om verschillende bacteriële soorten van elkaar te onderscheiden. Daarnaast hebben wij onderzocht welke parameters een indicatie voor een structurele of functionele rol kunnen geven. De keuze voor een lengte van 12 nucleotiden zorgde ervoor dat er voldoende zich herhalende sequenties gevonden konden worden, maar ook dat de kans op willekeurige herhaling relatief klein was. In principe kan het aantal geïdentificeerde potentieel functionele sequenties nog worden vergroot door tijdens het zoeken verschillende lengtes en een zekere mate van degeneratie toe te staan.

Om nieuw geïdentificeerde sequentie elementen te karakteriseren hebben wij een algemeen toepasbare strategie geformuleerd. In deze strategie wordt een distributie profiel gecreëerd voor elk zich herhalende sequentie die het volgende omschrijft: i) het aantal en de verspreiding over het genoom, ii) de verspreiding binnen de taxonomie, en iii) de locatie ten opzichte van de omliggende genen. In **hoofdstuk 3** worden verschillende herhaalde sequenties gekarakteriseerd en wordt de toepassing van de strategie toegelicht.

Er werden in totaal 583 oververtegenwoordigde sequenties gevonden met verschillende en soms intrigerende verspreidings-profielen, die potentieel waardevol zijn voor het identificeren van specifieke soorten of stammen. Verscheidene van deze sequenties waren al eerder beschreven, maar velen ook niet. Daarnaast was in veel gevallen geen biologische rol bekend. De meest voorkomende en breed verspreide sequenties in onze studie waren rijk aan Adenine. De Adenine-rijke herhaalde sequenties bleken aanwezig in redelijk grote aantallen per genoom en waren gelijkmatig verdeeld binnen de genomen waarin ze geïdentificeerd werden. Hoewel ze voornamelijk gevonden werden in intragene regios bleek er geen positionele voorkeur te zijn aangaande de lokale gen organisatie. Een dergelijk distributie profiel duidt op een globale rol, met een functie die niet gerelateerd is aan gen organisatie. Tegelijkertijd vonden we sequenties die heel specifiek waren voor verschillende soorten en stammen. Bij voorbeeld, de herhaalde sequentie ATGCCGTCTGAA werd zeer specifiek gevonden binnen het genus *Neisseria.*

Door gebruik te maken van een combinatie van de verschillende perspectieven op de verdeling en voorkomen van een sequentie, dat wil zeggen een breed taxomisch perspectief, een perspectief op genoom niveau en een perspectief gericht op de lokale gen organisatie, waren we in staat om een omschrijving van de geïdentificeerde sequenties te leveren. We nemen aan dat een vergelijkbare organisatie kan duiden op een vergelijkbare rol. Onze veralgemeende strategie kan dienen als een belangrijk fundament voor verder onderzoek aan de functie van dit type zich herhalende sequenties. Daarnaast kunnen de verkregen distributie profielen gebruikt worden om sequenties met vergelijkbare karakteristieken te vinden, waarvan de functie al bekend is. Vice versa laten we zien dat het distributie profiel gebruikt kan worden om de lijst met potentiële functionaliteiten in te perken. We hebben niet gepoogd een gedetailleerde functie aan alle gevonden sequenties toe te wijzen, omdat dit vele aanvullende analyses zou vereisen, die buiten het doel van dit hoofdstuk liggen. In plaats daarvan hebben we ervoor gekozen om een al eerder geïdentificeerde sequentie, met onbekende functie, uit *E. coli* met behulp van de beschreven strategie te analyseren. In **hoofdstuk 4** wordt dit werk beschreven.

Repetitieve Extragene Palindromische sequenties hebben een rol in het reduceren transcriptie interferentie in E. coli

Eén van de, in hoofdstuk 3 geïdentificeerde, zich herhalende sequenties met een intrigerend verspreidings-profiel bleken de zogenaamde 'Repetitieve Extragene Palindromische elementen'. Deze REPs zijn korte palindromische sequenties die in grote aantallen aanwezig zijn in de genomen van meerdere *Enterobacteriën*, waar ze tot wel 1% van het totale genoom bezetten. In het verleden zijn verschillende rollen aan deze REPs toegewezen, maar geen van deze rollen is verklarend voor alle gevonden REPs.

In **hoofdstuk 4** hebben we ons daarom gericht op het blootleggen van de primaire functie van de REPs. We zijn in deze opzet geslaagd door het analyseren van hun verspreiding over het genoom van het modelorganisme *E. coli* K12 MG1655 en het analyseren van de sequentie eigenschappen en expressie dynamiek van de omliggende genen. De REPs bevatten palindromische elementen en kunnen daardoor cruciforme structuren vormen. In het genoom van *E. coli* K12 MG1655 worden REPs bijna zonder uitzondering gevonden tussen genen met dezelfde oriëntatie (d.w.z. '→ REP →') en gen-paren met een convergente oriëntatie (d.w.z. '→ REP ←'). Deze gen organisaties zijn gevoelig voor de negatieve effecten van transcriptie ge-induceerde supercoiling (verlaagde expressie van downstream genen in een operon of van convergente genen). We vonden dat REPs voornamelijk

dichtbij het 3' einde van de geassocieerde genen gepositioneerd zijn. We vonden dat de upstream leden van gelijk georiënteerde gen-paren (d.w.z. '→ REP →') een significant hogere Codon Adaptation Index (CAI) hadden dan hun downstream leden (d.w.z. '→ REP →'). De CAI-waarden van convergente gen-paren gescheiden door een REP bleken ook significant hoger dan van hun tegenhangers zonder REP (d.w.z. '→ REP ←' versus '→ ←'). Deze duidelijke verschillen waren ook zichtbaar wanneer we gebruik maakten van daadwerkelijke gen-expressie data uit 466 microarray experimenten. Bij analyse van deze dataset bleek dat het gemiddelde verschil in expressie tussen convergente gen-paren met een REP kleiner was dan het verschil tussen de convergente gen-paren zonder REP, voor elk experiment binnen de dataset, terwijl de gemiddelde expressie zelfs hoger was. gepositioneerd zijn De bovenstaande observaties wezen in de richting van een coherente hypothese omtrent de rol van REPs. We concluderen dat REPs de negatieve effecten van transcriptie geïnduceerde supercoiling kunnen reduceren. REPs onderdrukken de supercoiling druk die ontstaat tijdens transcriptie door de formatie van cruciform structuren, waarmee ze een hogere gelijktijdige expressie van convergente gen-paren en van genen downstream van hoog tot expressie komende genen toestaan. Gegeven de spreiding van REP-achtige elementen in de soorten behorende tot de *gammaproteobacteria* is het aannemelijk dat dit mechanisme ook een rol speelt in andere soorten dan de *Enterobacteriën*.

Twee nieuwe methoden voor het evalueren van de kwaliteit van annotaties van translatie start posities

Het identificeren van coderende sequenties is niet alleen de eerste stap van een genoom-annotatie, het is ook een essentiële stap voor veel van de opvolgende analyses. Verschillende typen fouten kunnen gemaakt worden tijdens het voorspellen van coderende sequenties. Echte genen kunnen worden gemist (fout negatief) en voorspelde genen kunnen betekenisloos blijken (fout positief). Maar het meest uitdagende aspect van de voorspelling is de identificatie van de correcte start positie, ook wel TIS genoemd (d.w.z. translation initiation site). Incorrecte TIS-annotaties kunnen de identificatie van regulatoire elementen en signaal-peptiden schaden en kunnen sequentie-vergelijkingen negatief beïnvloeden. Gelukkig zijn er tegenwoordig verschillende computationele methoden, zoals Prodigal en Glimmer, die de coderende sequenties kunnen voorspellen met een relatief lage fout frequentie. Desalniettemin blijft de identificatie van coderende sequenties in genomen met een laag of hoog GC-gehalte voor deze methoden ook een uitdaging. Daarnaast bevatten de beschikbare publieke bronnen een substantieel aantal annotaties van mindere kwaliteit.

Om dit probleem aan te pakken hebben wij een referentie vrije methode ontwikkeld (eerste deel **hoofdstuk 5**) waarmee we in staat zijn om de kwaliteit van TIS-annotaties binnen een genoom te beoordelen. We hebben gevonden dat de correlatie tussen a) de geobserveerde distributie van alle potentiële TISs in de directe nabijheid van alle voorspelde coderende regio's en b) de kans-gebaseerde distributie van potentiële TISs, kan dienen als kwaliteits-maat voor de TIS annotatie. We hebben deze aanpak gevalideerd door de TIS-annotatie kwaliteit van verschillende model organismen, waarvan bekend is dat ze uitstekende TIS-annotaties hebben, te scoren. Eenmaal gevalideerd stelde deze methode ons in staat om de TIS-annotatie kwaliteit van alle publieke genomen te evalueren. We hebben daarbij vastgesteld dat de TIS-annotatie van de meerderheid van de genomen (87%) van voldoende kwaliteit lijkt, maar ook dat de TIS annotaties van 13% van de genomen sterk verbeterd zouden kunnen worden. We hebben onderzocht welke factoren een effect hebben op de TIS-annotatie kwaliteit. Van de onderzochte factoren bleek alleen het GC-gehalte van het geanalyseerde genoom een significant onderscheidend effect te hebben op de kwaliteit. De toenemende afhankelijkheid van computationele methoden heeft, in tegenstelling tot wat breed verondersteld werd, niet geresulteerd in een afname van annotatie kwaliteit over de jaren.

In het tweede gedeelte van **hoofdstuk 5** hebben we een strategie beschreven om huidige TIS-annotaties te evalueren en waar mogelijk, te verbeteren. Deze strategie maakt gebruik van een iteratieve PCA en is onafhankelijk van referentie data. Door middel van een simpel beoordelingsschema, wat de scores in de PCA-iteraties toepast, hebben wij de annotaties van verschillende genomen aangepast. Dit simpele beoordelingsschema bleek voldoende om de annotatie kwaliteit van veel van de slecht geannoteerde genomen in onze test-set te verbeteren. We hebben de kwaliteit van onze aanpassingen vergeleken met aanpassingen door Prodigal en vonden dat het genoom afhankelijk was welke methode de meeste verbeteringen bracht. Met behulp van een analyse aan het model organisme genoom van *E. coli* K12 MG1655 vonden we dat onze op PCA-gebaseerde strategie op grond van slechts 1 parameter de kwaliteit van de referentie methode Prodigal benaderde. De PCA-gebaseerde strategie kan daarom gebruikt worden om de uitkomst van bestaande methoden te evalueren en specifieke annotaties die mogelijk handmatige correcties vereisen markeren.

Slotopmerkingen

Het onderzoek beschreven in dit proefschrift richt zich op niet-coderende sequentie elementen met een (potentieel) regulerende rol. We hebben het bio-informatica onderzoek naar deze elementen gefaciliteerd door visualisatie tools te ontwikkelen met specifieke functionaliteiten en door de TIS-annotaties te beoordelen en te verbeteren. Onze strategie om sequentie elementen te karakteriseren doormiddel van locatie en taxonomie kan van grote waarde zijn in het identificeren en selecteren van sequenties met een potentieel regulerende rol. Het onderzoek beschreven in hoofdstuk 4 toont aan dat de analyses van onbekende sequentie elementen aanzienlijk versterkt kan worden door gebruik te maken van hoge kwaliteit annotaties en geïntegreerde genoom visualisaties. Het identificeren van een functionele rol voor deze ongekarakteriseerde sequentie elementen zal ons helpen het complexe pad van genotype naar fenotype beter te begrijpen.

Referenties

Ahlstrom, C., Barkema, H.W., and De Buck, J. (2016). Relative frequency of 4 major strain types of *Mycobacterium avium* ssp. paratuberculosis in Canadian dairy herds using a novel single nucleotide polymorphism-based polymerase chain reaction. *J. Dairy Sci*. **99**, 8297–8303.

Choi, S.C. (2016). On the study of microbial transcriptomes using second- and third-generation sequencing technologies. *J. Microbiol*. **54**, 527–536.

Ejby, M., Fredslund, F., Andersen, J.M., Vujičić Žagar, A., Henriksen, J.R., Andersen, T.L., Svensson, B., Slotboom, D.J., and Abou Hachem, M. (2016). An ATP Binding Cassette Transporter Mediates the Uptake of α-(1,6)-Linked Dietary Oligosaccharides in *Bifidobacterium* and Correlates with Competitive Growth on These Substrates. *J. Biol. Chem.* **291**, 20220–20231.

van den Esker, M.H., Kovács, Á.T., and Kuipers, O.P. (2016). YsbA and LytST are essential for pyruvate utilization in *Bacillus subtilis*. *Environ. Microbiol*.

Francke, C., Groot Kormelink, T., Hagemeijer, Y., Overmars, L., Sluijter, V., Moezelaar, R., and Siezen, R.J. (2011). Comparative analyses imply that the enigmatic sigma factor 54 is a central controller of the bacterial exterior. *BMC Genomics* **12**, 385.

Gennaris, A., Ezraty, B., Henry, C., Agrebi, R., Vergnes, A., Oheix, E., Bos, J., Leverrier, P., Espinosa, L., Szewczyk, J., et al. (2015). Repairing oxidized proteins in the bacterial envelope using respiratory chain electrons. *Nature* **528**, 409–412.

Greening, C., Biswas, A., Carere, C.R., Jackson, C.J., Taylor, M.C., Stott, M.B., Cook, G.M., and Morales, S.E. (2016). Genomic and metagenomic surveys of hydrogenase distribution indicate H2 is a widely utilised energy source for microbial growth and survival. *ISME J.* **10**, 761–777.

Houdt, R.V., and Mergeay, M. (2015). Genomic Context of Metal Response Genes in Cupriavidus metallidurans with a Focus on Strain CH34. In *Metal Response in Cupriavidus Metallidurans,* M. Mergeay, and R.V. Houdt, eds. (Springer International Publishing), pp. 21–44.

Khayatt, B.I., Overmars, L., Siezen, R.J., and Francke, C. (2013). Classification of the adenylation and acyltransferase activity of NRPS and PKS systems using ensembles of substrate specific hidden Markov models. *PloS One* **8**, e62136.

Lagares, A., Roux, I., and Valverde, C. (2016). Phylogenetic distribution and evolutionary pattern of an α-proteobacterial small RNA gene that controls polyhydroxybutyrate accumulation in *Sinorhizobium meliloti. Mol. Phylogenet. Evol.* **99**, 182–193.

Lee, I.-C., van Swam, I.I., Tomita, S., Morsomme, P., Rolain, T., Hols, P., Kleerebezem, M., and Bron, P.A. (2014). GtfA and GtfB are both required for protein O-glycosylation in *Lactobacillus plantarum*. *J. Bacteriol.* **196**, 1671–1682.

Liu, M., Prakash, C., Nauta, A., Siezen, R.J., and Francke, C. (2012). A computational analysis of cysteine and methionine metabolism and its regulation in dairy starter and related bacteria. *J. Bacteriol*. JB.06816-11.

Mehla, K., and Ramana, J. (2016). Identification of epitope-based peptide vaccine candidates against enterotoxigenic *Escherichia coli*: a comparative genomics and immunoinformatics approach. *Mol. Biosyst*. **12**, 890–901.

Ney, B., Ahmed, F.H., Carere, C.R., Biswas, A., Warden, A.C., Morales, S.E., Pandey, G., Watt, S.J., Oakeshott, J.G., Taylor, M.C., et al. (2016). The methanogenic redox cofactor F420 is widely synthesized by aerobic soil bacteria. *ISME J.*

Rauch, B.J., Gustafson, A., and Perona, J.J. (2014). Novel proteins for homocysteine biosynthesis in anaerobic microorganisms. *Mol. Microbiol.* **94**, 1330–1342.

Roche, B., Agrebi, R., Huguenot, A., Ollagnier de Choudens, S., Barras, F., and Py, B. (2015). Turning *Escherichia coli* into a Frataxin-Dependent Organism. *PLoS Genet.* **11**, e1005134.

Sghaier, H., Hezbri, K., Ghodhbane-Gtari, F., Pujic, P., Sen, A., Daffonchio, D., Boudabous, A., Tisa, L.S., Klenk, H.-P., Armengaud, J., et al. (2016). Stone-dwelling actinobacteria *Blastococcus saxobsidens, Modestobacter marinus* and *Geodermatophilus obscurus* proteogenomes. *ISME J.* **10**, 21–29.

Sheibani-Tezerji, R., Rattei, T., Sessitsch, A., Trognitz, F., and Mitter, B. (2015). Transcriptome Profiling of the Endophyte *Burkholderia phytofirmans* PsJN Indicates Sensing of the Plant Environment and Drought Stress. *mBio* **6**, e00621-615.

Siezen, R.J., Francke, C., Renckens, B., Boekhorst, J., Wels, M., Kleerebezem, M., and van Hijum, S.A.F.T. (2012). Complete resequencing and reannotation of the *Lactobacillus plantarum* WCFS1 genome. *J. Bacteriol.* **194**, 195–196.

Sorci, L., Ruggieri, S., and Raffaelli, N. (2014). NAD homeostasis in the bacterial response to DNA/RNA damage. *DNA Repair* **23**, 17–26.

Xiao, Y., van Hijum, S.A.F.T., Abee, T., and Wells-Bennik, M.H.J. (2015). Genome-Wide Transcriptional Profiling of *Clostridium perfringens* SM101 during Sporulation Extends the Core of Putative Sporulation Genes and Genes Determining Spore Properties and Germination Characteristics. *PloS One* **10**, e0127036.

ADDENDUM II

Dankwoord

Toen ik mijn PhD traject begon kreeg ik het idee om tegelijkertijd een opknapper van een huis te kopen en compleet te renoveren. Ik ben erachter gekomen dat het verbouwen van een huis verrassend veel overeenkomsten heeft met het voltooien van een proefschrift. Kamer voor kamer, hoofdstuk voor hoofdstuk ben ik verder gekomen. Elke klus leidt tot een nieuwe klus, waar elke analyse weer leidt tot een nieuwe analyse. En soms is het even doorbijten, als bijvoorbeeld de bestaande vloer ongeschikt blijkt en alles tot aan de kruipruimte gestript moet worden, of als voor de zoveelste keer die tijdrovende analyse over gedaan moet worden. Feit is dat mijn proefschrift inmiddels is afgerond en ik graag iedereen zou willen bedanken die heeft bijgedragen aan de totstandkoming.

Als eerste wil ik mijn promotor en co-promotoren bedanken. Christof, als co-promotor en directe begeleider ben jij een echte mentor voor mij geweest. Ik heb enorm veel van je geleerd. We hebben wetenschappelijk gezien misschien niet altijd voor de makkelijkste weg gekozen en bewust wat open deuren aan ons voorbij laten gaan. Maar juist door onze gezamenlijke zoektocht naar een alternatief perspectief durf ik mijzelf nu wetenschapper te noemen.

Roland, ik begon met mijn promotie in de periode dat jij emeritus hoogleraar werd. Bedankt dat je nu alsnog mijn promotor wil zijn. Je hebt me meegenomen als bezoekend onderzoeker op de UC Davis in de Verenigde Staten. Naast dat dit een hele leerzame periode was hebben we ook de tijd gehad om wijnen uit de Napa Valley te kunnen proeven en wist je me te verrassen met een voorliefde voor de Kentucky Fried Chicken.

Sacha, je bent in een later stadium betrokken bij mijn onderzoek, maar dat maakt mij niet minder dankbaar. Naast jouw eerlijke en nuchtere advies heb ik onze gesprekken over carrières en wetenschap altijd enorm kunnen waarderen.

Hoewel ik al inmiddels alweer even weg ben wil ik mijn (ex-) collega's van het CMBI en in het bijzonder de leden van de Bacterial Genomics groep bedanken. Specifiek wil ik even noemen Lennart, Victor, Juma, Tilman, Tom, Tom, (met drie Toms en een TomTom naar een congres in Luxemburg rijden), Fredrick, Bernadet, Michiel, Jos en Barbara. Verder heb ik het genoegen gehad om een aantal studenten te mogen begeleiden; Vincent, Brian, Yanick, Martijn, Daniel en Vincent v. D., bedankt voor jullie inzet! Daarnaast wil ik professor Gert Vriend noemen, die als hoofd van de afdeling altijd klaar staat voor zijn mensen.

Mijn recentere ex-collega's: Cherel, Tom, Catarina, Anca, Charlotte, Veerle, Muhe, Jason en Estelle, thanks for all the fun times at the UvA. Gerard, bedankt voor de fijne samenwerking, hoop dat er nog een aantal mooie artikelen uitrollen.

Heren van de schattige kniestukjes; Tim (denk dat we inmiddels met zekerheid mogen concluderen dat je geen hondsdolheid hebt opgelopen?), Joep, Tom v d B, Tom E, Rob, Eugene en Martin. Ik moet jullie teleurstellen, maar de grappen over mijn niet bestaande proefschrift kunnen vanaf vandaag echt niet meer. Gelukkig is mijn tuin nog steeds niet af en heb ik vertrouwen in jullie creativiteit. Heb zin in onze volgende trip naar een premium C-locatie in de wereld!

Paranimfen van dienst: Tom en Sander. Heb jullie beiden leren kennen tijdens de studies die tot dit proefschrift hebben geleid, zij het op verschillende momenten. Mooi dat ik dit nu met jullie beiden kan afsluiten! En achter elke succesvolle paranimf staat een sterke vrouw. Sophia en Marion, bij deze bedankt

Mijn moeder, Sonja. Je zal blij en trots zijn dat dit proefschrift er nu (eindelijk-) is. Als geen ander ken jij alle ups en downs die er aan vast hebben gezeten. Bedankt dat ik deze altijd met je heb kunnen delen. Het is enorm fijn om je altijd gesteund te voelen. Mijn vader, je bent een belangrijke pilaar geweest in dat andere grote project, mijn verbouwing. Daarnaast vrees ik toch echt dat ik jouw humor heb, bedankt. Peter, thanks bro en Karen, dank voor het in toom houden van mijn broertje.

En als allerlaatste, Emily. Ik geniet van elk moment dat we samen hebben. Bedankt voor alle steun en geduld. Op naar een mooie toekomst samen!

ADDENDUM III

About the author

Lex Overmars was born on 12th March, 1985 in Terwolde, the Netherlands. He obtained his MSc degree in Bioinformatics in 2009 from the Wageningen University. His master thesis entitled "Storage, analysis and integration of fermentation data, meta-data and ~omics data" was completed at Nizo Food Research, under the supervision of prof. Michiel Kleerebezem and dr. Michiel Wels. During his BSc studies, he co-founded Rijschoolvergelijker, the largest driving school comparison platform of the Netherlands. He conducted the research described in this thesis at the Center for Molecular and Biomolecular Informatics (CMBI), which is part of the Radboud Institute for Molecular Life Sciences (RIMLS) from the Radboudumc. After his PhD research, he worked as a Researcher at the University of Amsterdam. Since the first of February of this year, he is employed as data scientist at Vivat Insurances, which owns the brands Reaal and Zwitserleven.

ADDENDUM IV

List of publications

Overmars, L., Siezen, R. J., & Francke, C. (2015). A Novel Quality Measure and Correction Procedure for the Annotation of Microbial Translation Initiation Sites. *PloS One*, **10**(7), e0133691.

Overmars, L., van Hijum, S. A. F. T., Siezen, R. J., & Francke, C. (2015). CiVi: circular genome visualization with unique features to analyze sequence elements. *Bioinformatics*, **31**(17), 2867–2869.

Ederveen, T. H. A., **Overmars, L.**, & van Hijum, S. A. F. T. (2013). Reduce manual curation by combining gene predictions from multiple annotation engines, a case study of start codon prediction. *PLoS One*, **8**(5), e63523.

Khayatt, B. I., **Overmars, L.**, Siezen, R. J., & Francke, C. (2013). Classification of the adenylation and acyl-transferase activity of NRPS and PKS systems using ensembles of substrate specific hidden Markov models. *PLoS One*, **8**(4), e62136.

Overmars, L., Kerkhoven, R., Siezen, R. J., & Francke, C. (2013). MGcV: the microbial genomic context viewer for comparative genome analysis. *BMC Genomics*, **14**, 209.

Bron, P. A., Wels, M., Bongers, R. S., van Bokhorst-van de Veen, H., Wiersma, A., **Overmars, L.**, … Kleerebezem, M. (2012). Transcriptomes reveal genetic signatures underlying physiological variations imposed by different fermentation conditions in *Lactobacillus plantarum. PLoS One*, **7**(7), e38720.

Kormelink, T. G., Koenders, E., Hagemeijer, Y., **Overmars, L.**, Siezen, R. J., de Vos, W. M., & Francke, C. (2012). Comparative genome analysis of central nitrogen metabolism and its control by GlnR in the class Bacilli. *BMC Genomics*, **13**, 191.

Touw, W. G., Bayjanov, J. R., **Overmars, L.**, Backus, L., Boekhorst, J., Wels, M., & van Hijum, S. A. F. T. (2012). Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle? *Brief. Bioinform.* **14**(3), 315

Francke, C., Kormelink, T. G., Hagemeijer, Y., **Overmars, L.**, Sluijter, V., Moezelaar, R., & Siezen, R. J. (2011). Comparative analyses imply that the enigmatic Sigma factor 54 is a central controller of the bacterial exterior. *BMC Genomics*, **12**, 385.

Wels, M., **Overmars, L.**, Francke, C., Kleerebezem, M., & Siezen, R. J. (2011). Reconstruction of the regulatory network of *Lactobacillus plantarum* WCFS1 on basis of correlated gene expression and conserved regulatory motifs. *Microbial Biotechnology*, *4*(3), 333–344.